nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-022-01410-x

Check for updates

Acoustic regularities in infant-directed speech and song across cultures

Courtney B. Hilton[1,2,44] ✉, Cody J. Moser[1,3,44] ✉, Mila Bertolo[1], Harry Lee-Rubin[1], Dorsa Amir[4], Constance M. Bainbridge[1,5], Jan Simson[1,6], Dean Knox[7], Luke Glowacki[8], Elias Alemu[9], Andrzej Galbarczyk[10], Grazyna Jasienska[10], Cody T. Ross[11], Mary Beth Neff[12,13], Alia Martin[12], Laura K. Cirelli[14,15], Sandra E. Trehub[15], Jinqi Song[16], Minju Kim[17], Adena Schachner[17], Tom A. Vardy[18], Quentin D. Atkinson[18,19], Amanda Salenius[20], Jannik Andelin[20], Jan Antfolk[20], Purnima Madhivanan[21,22,23,24], Anand Siddaiah[24], Caitlyn D. Placek[25], Gul Deniz Salali[26], Sarai Keestra[26,27], Manvir Singh[28,29], Scott A. Collins[30], John Q. Patton[31], Camila Scaff[32], Jonathan Stieglitz[29,33], Silvia Ccari Cutipa[34], Cristina Moya[35,36], Rohan R. Sagar[37,38], Mariamu Anyawire[39], Audax Mabulla[40], Brian M. Wood[41], Max M. Krasnow[1,42] and Samuel A. Mehr[1,2,43] ✉

When interacting with infants, humans often alter their speech and song in ways thought to support communication. Theories of human child-rearing, informed by data on vocal signalling across species, predict that such alterations should appear globally. Here, we show acoustic differences between infant-directed and adult-directed vocalizations across cultures. We collected 1,615 recordings of infant- and adult-directed speech and song produced by 410 people in 21 urban, rural and small-scale societies. Infant-directedness was reliably classified from acoustic features only, with acoustic profiles of infant-directedness differing across language and music but in consistent fashions. We then studied listener sensitivity to these acoustic features. We played the recordings to 51,065 people from 187 countries, recruited via an English-language website, who guessed whether each vocalization was infant-directed. Their intuitions were more accurate than chance, predictable in part by common sets of acoustic features and robust to the effects of linguistic relatedness between vocalizer and listener. These findings inform hypotheses of the psychological functions and evolution of human communication.

The forms of many animal signals are shaped by their functions, a link arising from production- and reception-related rules that help to maintain reliable signal detection within and across species[1–6]. Form–function links are widespread in vocal signals across taxa, from meerkats to fish[3,7–10], causing acoustic regularities that allow cross-species intelligibility[11–14]. This facilitates the ability of some species to eavesdrop on the vocalizations of other species, for example, as in superb fairywrens (*Malurus cyaneus*), who learn to flee predatory birds in response to alarm calls that they themselves do not produce[15].

In humans, an important context for the effective transmission of vocal signals is between parents and infants, as human infants are particularly helpless[16]. To elicit care, infants use a distinctive alarm signal: they cry[17]. In response, adults produce infant-directed language and music (sometimes called 'parentese') in forms of speech and song with putatively stereotyped acoustics[18–35].

These stereotyped acoustics are thought to be functional: supporting language acquisition[36–39], modulating infant affect and temperament[33,40,41] and/or coordinating communicative interactions with infants[42–44]. These theories all share a key prediction: like the vocal signals of other species, the forms of infant-directed vocalizations should be shaped by their functions, instantiated with clear regularities across cultures. Put another way, we should expect people to alter the acoustics of their vocalizations when those vocalizations are directed toward infants and they should make those alterations in similar fashions worldwide.

The evidentiary basis for such a claim is controversial, however, given the limited generalizability of individual ethnographic reports and laboratory studies[45], small stimulus sets[46] and a variety of counter-examples[47–53]. Some evidence suggests that infant-directed speech is primarily characterized by higher and more variable pitch[54] and more exaggerated and variable vowels[23,55,56], on the basis of many studies in modern industrialized societies[23,28,57–61] and a few in small-scale societies[62,63]. Infants are themselves sensitive to these features, preferring them, even if spoken in unfamiliar languages[64–66]. But these acoustic features are less exaggerated or reportedly absent in some cultures[51,59,67] and may vary in relation to the age and sex of the infant[59,68,69], weighing against claims of cross-cultural regularities.

In music, infant-directed songs also seem to have some stereotyped acoustic features. Lullabies, for example, tend toward slower tempos, reduced accentuation and simple repetitive melodic patterns[31,32,35,70], supporting functional roles associated with infant care[33,41,42] in both industrialized[34,71–73] and small-scale societies[74,75]. Infants are soothed by these acoustic features, whether produced in familiar[76,77] or unfamiliar songs[78] and both adults and children reliably associate the same features with a soothing function[31,32,70]. But cross-cultural studies of infant-directed song have primarily relied

A full list of affiliations appears at the end of the paper.

upon archival recordings from disparate sources[29,31,32], an approach that poorly controls for differences in voices, behavioural contexts, recording equipment and historical conventions, limiting the precision of findings and complicating their generalizability.

Measurements of the same voices producing multiple vocalizations, gathered from many people in many languages, worldwide, would enable the clearest analyses of whether and how humans alter the acoustics of their vocalizations when communicating with infants, helping to address the lack of consensus in the literature. Further, yoked analyses of both speech and song may explain how the forms of infant-directed vocalizations reliably differ from one another, testing theories of their shared or separate functions[33,36–42].

We take this approach here. We built a corpus of infant-directed speech, adult-directed speech, infant-directed song and adult-directed song from 21 human societies, totalling 1,615 recordings of 410 voices (Fig. 1a, Table 1 and Methods; the corpus is open-access at https://doi.org/10.5281/zenodo.5525161). We aimed to maximize linguistic, cultural, geographic and technological diversity: the recordings document vocalizations in 18 languages from 12 language families and represent societies located on six continents, with varying degrees of isolation from global media, including four small-scale societies that lack access to television, radio or the internet and therefore have strongly limited exposure to language and music from other societies. Participants were asked to provide all four vocalization types.

We used computational analyses of the acoustic forms of the vocalizations and a citizen-science experiment to test (1) the degree to which infant-directed vocalizations are cross-culturally stereotyped and (2) the degree to which naive listeners detect infant-directedness in language and music.

Results

Infant-directed vocalization is cross-culturally stereotyped. We studied 15 types of acoustic features in each recording (for example, pitch, rhythm and timbre) via 94 summary variables (for example, median and interquartile range (IQR)) that were treated to reduce the influence of atypical observations, such as extreme values caused by loud wind, rain and other background noises (Methods and Supplementary Methods; a codebook is in Supplementary Table 1). To minimize the potential for bias, we collected the acoustic data using automated signal extraction tools that measure physical characteristics of the auditory signal; such physical characteristics lack cultural information (in contrast to, for example, human annotations) and thus can be applied reliably across diverse audio recordings.

First, we asked whether the acoustics of infant-directed speech and song are stereotyped in similar ways across the societies whose recordings we studied. Following previous work[32], we used a least absolute shrinkage and selection operator (LASSO) logistic classifier[79] with fieldsite-wise *k*-fold cross-validation, separately for speech and song recordings, using all 15 types of acoustic features (Methods). This approach provides a strong test of cross-cultural regularity: the model is trained only on data from 20 of the 21 societies to predict whether each vocalization in the twenty-first society is infant- or adult-directed. The procedure is repeated 20 further times, with each society being held out, ensuring the model is trained evenly across the full set of societies. The summary of the model's performance reflects, corpus-wide, the degree to which infant-directed speech and song are acoustically stereotyped, as high classification performance can only result from cross-cultural regularities.

The models accurately classified both speech and song, on average, across and within societies, with above-chance performance in 21 of 21 fieldsites for both speech and song (Fig. 1b; speech: area under the curve (AUC) = 91%, 95% confidence interval (CI) (86%, 96%); song: AUC = 82%, 95% CI (76%, 89%)).

To test the reliability of these findings, we repeated them with two alternate strategies, using the same cross-validation procedure but doing so across language families and geographic regions instead of fieldsites. The results robustly replicated in both cases (Supplementary Fig. 1). Moreover, to ensure that the main LASSO results were not attributable to particulars of the audio-editing process (Methods), we also repeated them using unedited audio from the corpus; the results replicated again (Supplementary Fig. 2).

These findings show that the acoustic features of infant-directed speech and song are robustly stereotyped across the 21 societies studied here.

Infant-directedness differs acoustically in speech and song. We used two convergent approaches to determine the specific acoustic features that are predictive of infant-directedness in speech and song.

First, the LASSO procedure identified the most reliable predictors of contrasts between infant- and adult-directed vocalizations. The most influential of these predictors are reported in Fig. 1b, with their relative variable importance scores. These show substantial differences in the variables the model relied upon to reliably classify speech and song across cultures. For example, pitch (F_0 median and IQR) and median vowel travel rate strongly differentiated infant-directedness in speech but not in song, while vowel travel variability (IQR) and median intensity strongly differentiated infant-directedness in song but not in speech. The full results of the LASSO variable selection are in Supplementary Table 2.

Second, in a separate exploratory–confirmatory analysis, we used mixed-effects regression to measure the expected difference in each acoustic feature associated with infant-directedness, separately for speech and song. Importantly, this approach estimates main effects adjusted for sampling variability and estimates fieldsite-level effects, allowing for tests of the degree to which the main effects differ in magnitude across cultures (for example, for a given acoustic feature, if recordings from some fieldsites show larger differences between infant- and adult-directed speech than do recordings from other fieldsites). The analysis was preregistered.

The procedure identified 11 acoustic features that reliably distinguished infant-directedness in song, speech or both (Fig. 2; statistics are in Supplementary Table 3); we also estimated these effects within each fieldsite (see the doughnut plots in Fig. 2 and full estimates in Extended Data Fig. 1).

In speech, across all or most societies, infant-directedness was characterized by higher pitch, greater pitch range and more contrasting vowels than was adult-directed speech from the same voices (largely replicating the results of the LASSO approach; Fig. 1b and Supplementary Table 2). Several acoustic effects were consistent in all fieldsites (for example, pitch, energy roll-off and pulse clarity), while other features, such as vowel contrasts and inharmonicity, were consistent in most of them. These patterns align with prior claims of pitch and vowel-contrast being robust features of infant-directed speech[23,60] and substantiate them across many cultures.

The distinguishing features of infant-directed song were more subtle than those of speech but nevertheless corroborate its purported soothing functions[33,41,42]: reduced intensity and acoustic roughness, although these were less consistent across fieldsites than the speech results. The less-consistent effects may result from the fact that, while solo-voice speaking is fairly natural and representative of most adult-directed speech (that is, people rarely speak at the same time), much of the world's song occurs in social groups where there are multiple singers and accompanying instruments[32,42,80]. Asking participants to produce solo adult-directed song may have biased participants toward choosing more soothing and intimate songs (for example, ballads, love songs; Supplementary Table 4) or less naturalistic renditions of songs. The production of songs in the presence of an infant may also have altered participants' singing

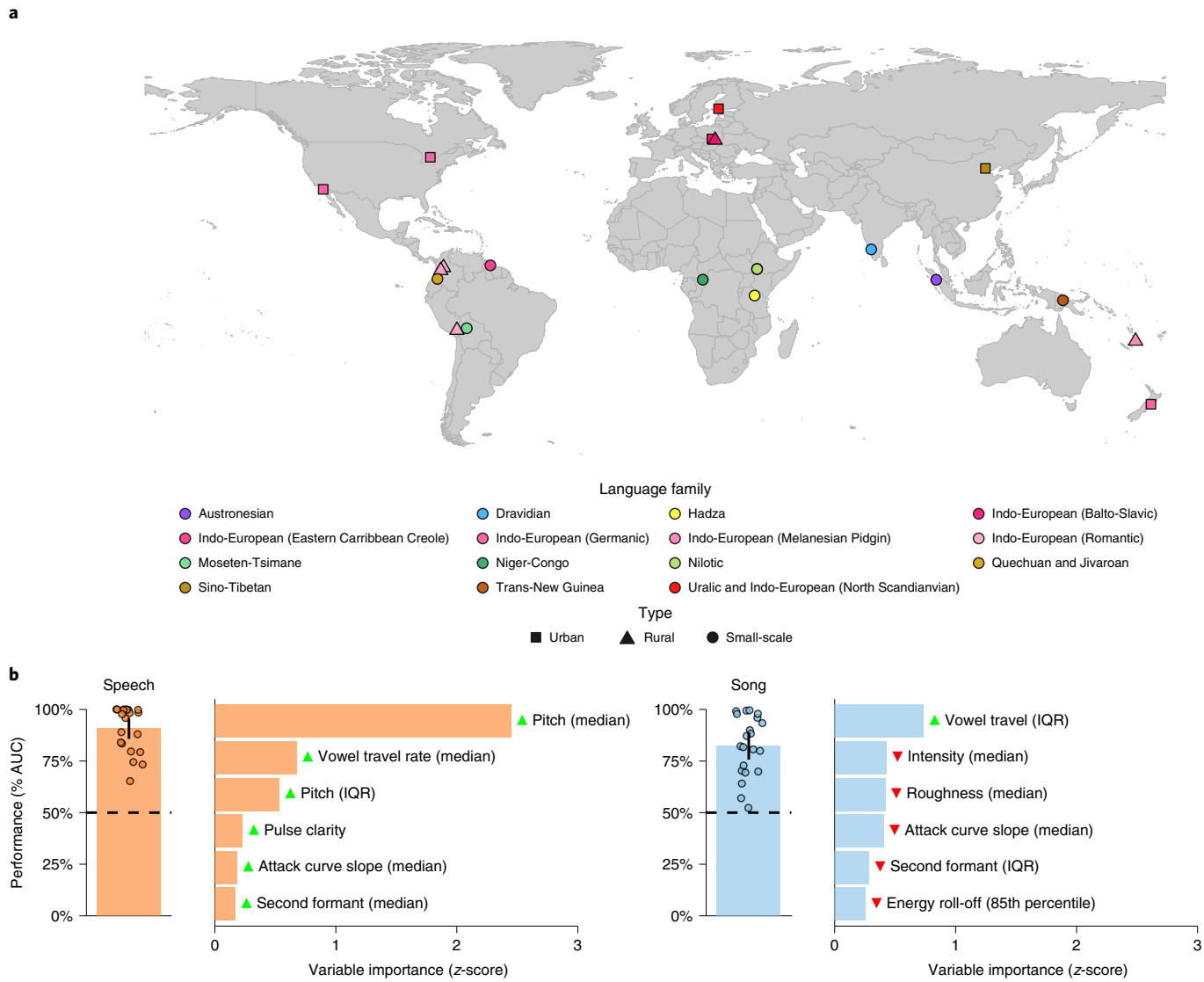


Fig. 1 | Cross-cultural regularities in infant-directed vocalizations. a, We recorded examples of speech and song from 21 urban, rural or small-scale societies, in many languages. The map indicates the approximate location of each society and is colour-coded by the language family or subgroup represented by the society. **b**, Machine-learning classification demonstrates the stereotyped acoustics of infant-directed speech and song. We trained two LASSO models, one for speech and one for song, to classify whether recordings were infant- or adult-directed on the basis of their acoustic features. These predictors were regularized using fieldsite-wise cross-validation, such that the model optimally classified infant-directedness across all 21 societies studied. The vertical bars represent the mean classification performance across societies ($n = 21$ societies for both speech and song; quantified via receiver operating characteristic/AUC); the error bars represent 95% CI of the mean; the points represent the performance estimate for each fieldsite; and the horizontal dashed lines indicate chance level of 50% AUC. The horizontal bars show the six acoustic features with the largest influence in each classifier; the green and red triangles indicate the direction of the effect, for example, with median pitch having a large, positive effect on classification of infant-directed speech. The full results of the variable selection procedure are in Supplementary Table 2, with further details in Methods.

style[35]. Thus, the distinctiveness of infant-directed song (relative to adult-directed song) may be underestimated here.

The exploratory–confirmatory analyses provided convergent evidence for opposing acoustic trends across infant-directed speech and song, as did an alternate approach using principal-components analysis (PCA); three principal components most strongly distinguished speech from song, infant-directed song from adult-directed song and infant-directed speech from adult-directed speech (Supplementary Results and Extended Data Fig. 2). Replicating the LASSO findings, for example, median pitch strongly differentiated infant-directed speech from adult-directed speech but it had no such effect in music; pitch variability had the opposite effect across language and music; and further differences were evident in pulse clarity, inharmonicity and energy roll-off. These patterns are consistent with the possibility of differentiated functional roles across infant-directed speech and song[18,33,34,42,77,78,81].

Some acoustic features were nevertheless common to both language and music. In particular, overall, infant-directedness was characterized by reduced roughness, which may facilitate parent–infant signalling[5,41] through better contrast with the sounds of screaming or crying[17,82]. Infant-directedness was also characterized by increased vowel contrasts, potentially to aid language acquisition[36,37,39] or as a byproduct of socio-emotional signalling[1,56].

Listeners are sensitive to infant-directedness. If people worldwide reliably alter their speech and song when interacting with infants,

Table 1 | Societies from which recordings were gathered

Region	Subregion	Society	Language	Language family	Subsistence type	Population	Distance to city (km)	Children per family	Recordings
Africa	Central Africa	Mbendjele BaYaka	Mbendjele	Niger-Congo	Hunter-gatherer	61–152	120	7	60
	Eastern Africa	Hadza	Hadza	Hadza	Hunter-gatherer	35	80	6	38
		Nyangatom	Nyangatom	Nilotic	Pastoralist	155	180	5.6	56
		Toposa	Toposa	Nilotic	Pastoralist	250	180	5.2	60
Asia	East Asia	Beijing	Mandarin	Sino-Tibetan	Urban	21.5 million	0	1	124
	South Asia	Jenu Kurubas	Kannada	Dravidian	Other	2,000	15	1	80
	Southeast Asia	Mentawai Islanders	Mentawai	Austronesian	Horticulturalist	260	120	Unknown	60
Europe	Eastern Europe	Krakow	Polish	Indo-European	Urban	771,069	0	1.54	44
		Rural Poland	Polish	Indo-European	Agriculturalists	6,720	70	1.83	55
	Scandinavia	Turku	Finnish and Swedish	Uralic and Indo-European	Urban	186,000	0	1.41	80
North America	North America	San Diego	English (United States)	Indo-European	Urban	3.3 million	0	1.7	116
		Toronto	English (Canadian)	Indo-European	Urban	5.9 million	0	1.5	198
Oceania	Melanesia	Ni-Vanuatu	Bislama	Indo-European Creole	Horticulturalist	6,000	224	3.78	90
		Enga	Enga	Trans-New Guinea	Horticulturalist	500	120	6	22
	Polynesia	Wellington	English (New Zealand)	Indo-European	Urban	210,400	0	1.45	228
South America	Amazonia	Arawak	English (Creole)	Indo-European	Other	350	32	3	48
		Tsimane	Tsimane	Moseten-Tsimane	Horticulturalist	150	234	9	51
		Sapara and Achuar	Quechua and Achuar	Quechuan and Jivaroan	Horticulturalist	200	205	9	59
	Central Andes	Quechua/Aymara	Spanish	Indo-European	Agro-pastoralist	200	8	4	49
	Northwestern South America	Afrocolombians	Spanish	Indo-European	Horticulturalist	300–1,000	100	6.6	53
		Colombian Mestizos	Spanish	Indo-European	Commercial economy	470,000	0	3.5	43

as the above findings demonstrate, this may enable listeners to make reliable inferences concerning the intended targets of speech and song, consistent with functional accounts of infant-directed vocalization[33,36–42,83,84]. We tested this secondary hypothesis in a simple listening experiment, conducted in English using web-based citizen-science methods[85].

We played excerpts from the vocalization corpus to 51,065 people (after exclusions; Methods) in the 'Who's Listening?' game on The Music Lab, a citizen-science platform for auditory research. The participants resided in 187 countries (Fig. 3b) and reported speaking 199 languages fluently (including second languages, for bilinguals). We asked them to judge, quickly, whether each vocalization was directed to a baby or to an adult (Methods and Extended Data Fig. 3). Readers may participate in the naive listener experiment by visiting https://themusiclab.org/quizzes/ids.

The responses were strongly biased toward 'baby' responses when hearing songs and away from 'baby' responses when hearing speech, regardless of the actual target of the vocalizations (Extended Data Fig. 4). To correct for these response biases, we used *d*-prime analyses at the level of each vocalist; that is, analysing listeners' sensitivity to infant-directedness in speech and song (Supplementary Methods). Unless noted otherwise, all estimates reported here are generated by mixed-effects linear regression, adjusting for fieldsite nested within world region, via random effects.

The listeners' intuitions were accurate, on average and across fieldsites (Fig. 3a; response times shown in Extended Data Fig. 5). Sensitivity (d') was significantly higher than the chance level of 0 (speech: $d'=1.19$, $t_{4.65}=3.63$, 95% CI (0.55, 1.83), $P=0.017$; song: $d'=0.51$, $t_{4.52}=3.06$, 95% CI (0.18, 0.83), $P=0.032$; note, all P values reported in this paper are two-sided). These results were robust to learning effects (Supplementary Fig. 3) and to multiple data trimming decisions. For example, they repeated whether or not recordings with confounding contextual/background cues (for example, an audible infant) were excluded and also when data from English-language recordings, which were probably understandable to participants, were excluded (Supplementary Results).

To test the consistency of listener inferences across cultures, we estimated fieldsite-level sensitivity from the random effects in the model. Cross-site variability was evident in the magnitude of sensitivity effects: listeners were far better at detecting infant-directedness in some sites than others (with high d' in recordings from Wellington, New Zealand, for both speech and song, but marginal d' in recordings from Tannese Vanuatans, for example). Nevertheless, the estimated mean fieldsite-wise d' was greater than 0 in both speech and song in all fieldsites (Fig. 3a) with 95% CI not overlapping with 0 in 18 of 21 field sites for speech and 16 of 20 for song (Supplementary Table 5; one d' estimate could not be computed for song due to missing data). Most fieldsite-wise sample sizes after exclusions were small (Methods), so we caution that fieldsite-wise estimates are far less interpretable than the overall d' estimate reported above.

Analyses of cross-cultural variability among listeners revealed similarities in their perception of infant-directedness. In particular, coefficient of variation scores revealed little variation in listener accuracy across countries of origin (2.3%) and native languages (1.1%), with the estimated effects of age and gender both less than 1%. And more detailed demographic characteristics available for a subset of participants in the United States, including socioeconomic status and ethnicity, also explained little variation in accuracy (Supplementary Results). These findings suggest general cross-demographic consistency in listener intuitions.

One important aspect of listeners was predictive of their performance, however: their degree of relatedness to the vocalizer, on a given trial. To analyse this, we estimated fixed effects for three forms of linguistic relatedness between listener and vocalizer: (1) weak relatedness, when a language the listener spoke fluently was from a different language family than that of the vocalization (for example, when the vocalization was in Mentawai, an Austronesian

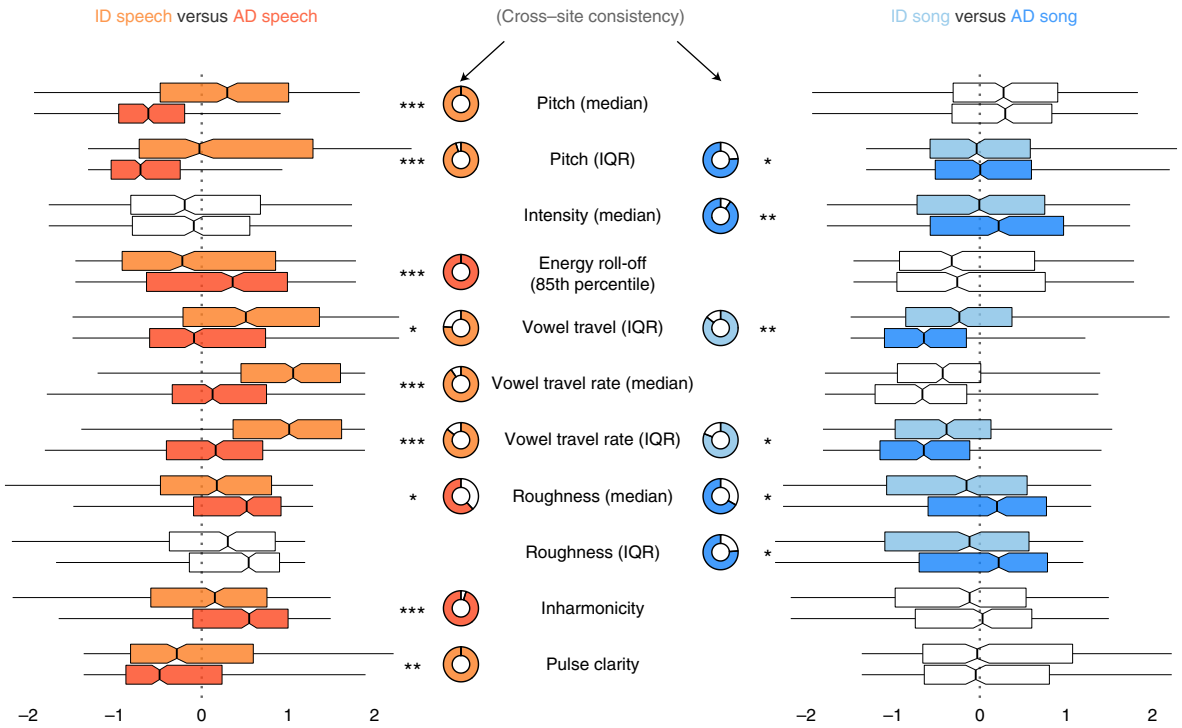


Fig. 2 | How people alter their voices when vocalizing to infants. Eleven acoustic features had a statistically significant difference between infant- and adult-directed vocalizations, within-voices, in speech, song or both. Consistent with the LASSO results (Fig. 1b and Supplementary Table 2), the acoustic features operated differently across speech and song. For example, median pitch was far higher in infant-directed speech than in adult-directed speech, whereas median pitch was comparable across both forms of song. Some features were highly consistent across fieldsites (for example, lower inharmonicity in infant-directed speech than adult-directed speech), whereas others were more variable (for example, lower roughness in infant-directed speech than in adult-directed speech). The boxplots, which are ordered approximately from largest to smallest differences between effects across speech and song, represent each acoustic feature's median (vertical black lines) and IQR (boxes); the whiskers indicate 1.5× IQR; the notches represent the 95% CI of the medians; and the doughnut plots represent the proportion of fieldsites where the main effect repeated, based on estimates of fieldsite-wise random effects. Only comparisons that survived an exploratory–confirmatory analysis procedure are plotted; the faded boxplots denote comparisons that did not reach statistical significance in confirmatory analyses. Significance values are computed via linear combinations with two-sided tests, following multilevel mixed-effects models ($n = 1{,}570$ recordings); $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$; no adjustments were made for multiple comparisons, given the exploratory–confirmatory approach taken. Regression results are in Supplementary Table 3 and a full report of fieldsite-level estimates is in Supplementary Table 5. Note: the model estimates are normalized jointly on speech and song data so as to enable comparisons across speech and song for each feature; as such, the absolute distance from 0 for a given feature is not directly interpretable but estimates are directly comparable across speech and song. ID, infant-directed; AD, adult-directed.

language, and the listener spoke fluent Mandarin, a Sino-Tibetan language); (2) moderate relatedness, when the languages were from the same language family (for example, when the vocalization was in Spanish and the listener spoke fluent English, which are both Indo-European languages); or (3) strong relatedness, when a language the listener spoke fluently exactly matched the language of the vocalization.

Sensitivity was significantly above chance in all cases (Fig. 3c), with increases in performance associated with increasing relatedness (unrelated: estimated speech $d' = 1.03$, song $d' = 0.37$; same language family: speech $d' = 1.31$, song $d' = 0.65$; same language: speech $d' = 1.58$, song $d' = 0.92$). Some of this variability is probably attributable to trivial language comprehensiblity; that is, in cases of strong relatedness, listeners very likely understood the words of the vocalization, strongly shaping their infant-directedness rating.

These findings provide an important control, as they demonstrate that the overall effects reported in the naive listener experiment (Fig. 3a) are not attributable to linguistic similarities between listeners and vocalizers (Fig. 3c), which could, for example, allow listeners to detect infant-directedness on the basis of the words or other linguistic features of the vocalizations, as opposed to their acoustic features. And while the instructions for the experiment were presented in English (suggesting that all listeners probably had at least a cursory understanding of English), the findings were robust to the exclusion of all English-language recordings (Supplementary Results).

We also found suggestive evidence of other, non-linguistic links between listeners and vocalizers being predictive of sensitivity. For example, fieldsite population size and distance to the nearest urban centre were correlated estimated sensitivity to infant-directedness in that fieldsite. These and similar effects (Supplementary Results) suggest that performance was somewhat higher in the larger, more industrialized fieldsites that are more similar to the environments of internet users, on average. But these analyses are necessarily coarser than the linguistic relatedness tests reported above.

Listener intuitions are modulated by vocalization acoustics. Last, we studied the degree to which the acoustic features of the recordings were predictive of listeners' intuitions concerning them (measured as the experiment-wide proportions of infant-directedness ratings for each vocalization, in a similar approach to other research[70]). These proportions can be considered a continuous measure of perceived infant-directedness, per the ears of the naive listeners. We trained two LASSO models to predict the proportions, with the

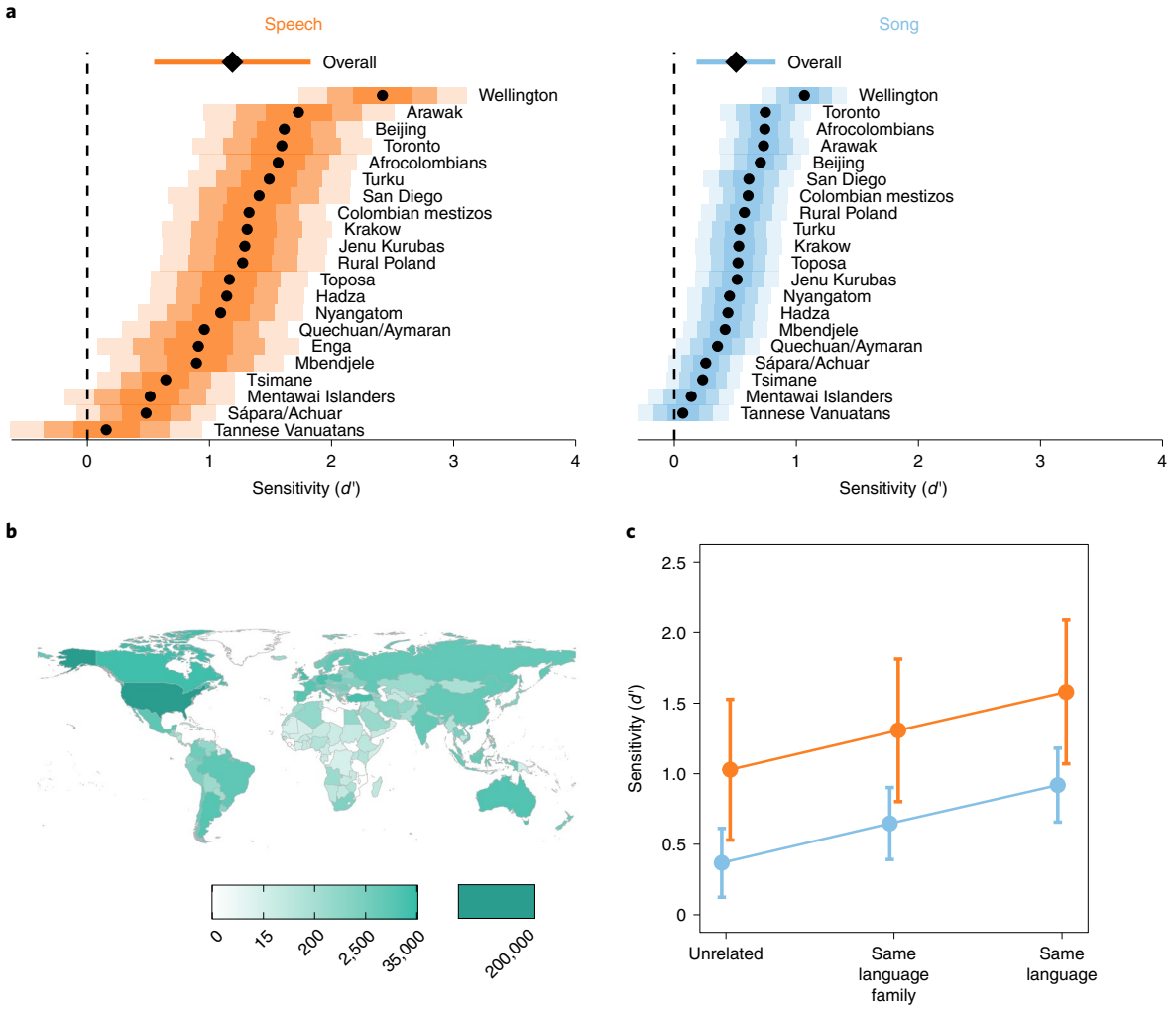


Fig. 3 | Naive listeners distinguish infant-directed vocalizations from adult-directed vocalizations across cultures. Participants listened to vocalizations drawn at random from the corpus, viewing the prompt 'Someone is speaking or singing. Who do you think they are singing or speaking to?' They could respond with either 'adult' or 'baby' (Extended Data Fig. 3). From these ratings (after exclusion $n = 473$ song recordings; $n = 394$ speech recordings), we computed listener sensitivity (d'). **a**, Listeners reliably detected infant-directedness in both speech and song, overall (indicated by the diamonds, with 95% CI indicated by the horizontal lines) and across many fieldsites (indicated by the black dots), although the strength of the fieldsite-wise effects varied substantially (see the distance between the vertical dashed line and the black dots; the shaded regions represent 50%, 80% and 95% CI, in increasing order of lightness). Note that one fieldsite-wise d' could not be estimated for song; complete statistical reporting is in Supplementary Table 5. **b**, The participants in the citizen-science experiment hailed from many countries; the gradients indicate the total number of vocalization ratings gathered from each country. **c**, The main effects held across different combinations of the linguistic backgrounds of vocalizer and listener. We split all trials from the main experiment into three groups: those where a language the listener spoke fluently was the same as the language of the vocalization ($n = 82{,}094$), those where a language the listener spoke fluently was in the same major language family as the language of the vocalization ($n = 110{,}664$) and those with neither type of relation ($n = 285{,}378$). The plot shows the estimated marginal effects of a mixed-effects model predicting d' values across language and music examples, after adjusting for fieldsite-level effects. The error bars represent 95% CI of the mean. In all three cases, the main effects replicated; increases in linguistic relatedness corresponded with increases in sensitivity.

same fieldsite-wise cross-validation procedure used in the acoustic analyses reported above. Both models explained variation in human listeners' intuitions, albeit more so in speech than in song (Fig. 4; speech $R^2 = 0.59$; song $R^2 = 0.18$, both $P < 0.0001$; P values calculated using robust standard errors), probably because the acoustic features studied here more weakly guided listeners' intuitions in song than they did in speech.

If human inferences are attuned to cross-culturally reliable acoustic correlates of infant-directedness, one might expect a close relationship between the strength of actual acoustic differences between vocalizations on a given feature and the relative influence of that feature on human intuitions. To test this question, we correlated how strongly a given acoustic feature distinguished infant-directed from adult-directed speech and song (Fig. 2; estimated with mixed-effects modelling) with the variable importance of that feature in the LASSO model trained to predict human intuitions (the barplots in Fig. 4). We found a strong positive relationship for speech ($r = 0.72$) and a weaker relationship for song ($r = 0.36$).

This difference may help to explain the weaker intuitions of the naive listeners in song, relative to speech: naive listeners' inferences about speech were more directly driven by acoustic features that actually characterize infant-directed speech worldwide, whereas their inferences about song were erroneously driven by acoustic features that less reliably characterize infant-directed song worldwide. For example, songs with higher pulse clarity and median second formants and lower median first formants were more likely to be rated

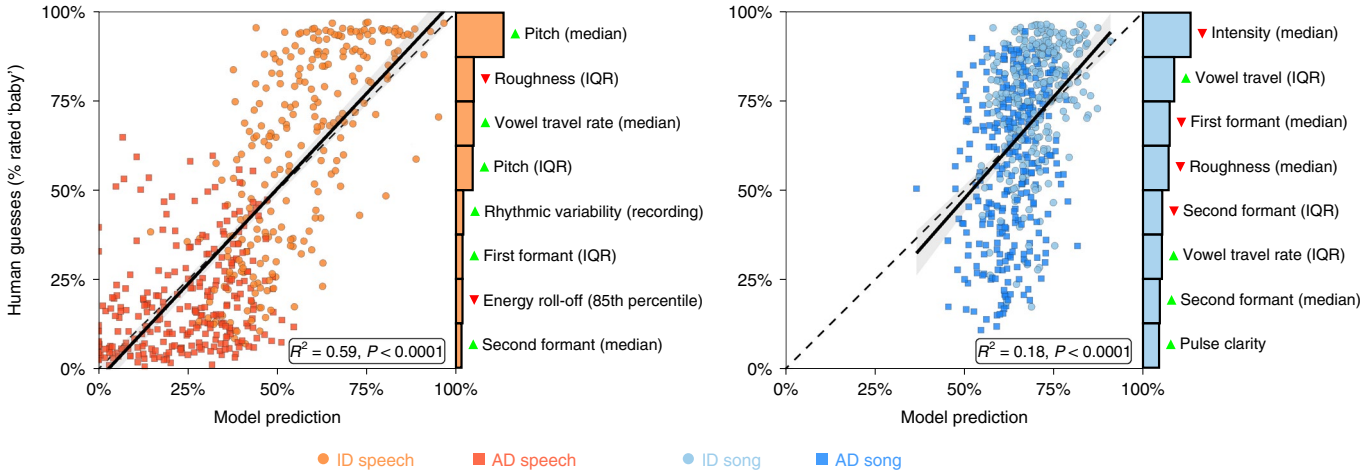


Fig. 4 | Human inferences about infant-directedness are predictable from acoustic features of vocalizations. To examine the degree to which human inferences were linked to the acoustic forms of the vocalizations, we trained two LASSO models to predict the proportion of 'baby' responses for each non-confounded recording from the human listeners. While both models explained substantial variability in human responses, the model for speech was more accurate than the model for song, in part because the human listeners erroneously relied on acoustic features for their predictions in song that less reliably characterized infant-directed song across cultures (Figs. 1b and 2). Each point represents a recorded vocalization (after exclusions $n = 528$ speech recordings; $n = 587$ song recordings), plotted in terms of the model's estimated infant-directedness of the model and the average 'infant-directed' rating from the naive listeners; the barplots depict the relative explanatory power of the top eight acoustical features in each LASSO model, showing which features were most strongly associated with human inferences (the green or red triangles indicate the directions of effects, with green higher in infant-directed vocalizations and red lower); the dashed diagonal lines represent a hypothetical perfect match between model predictions and human guesses; the solid black lines depict linear regressions (speech: $F(1,526) = 773$, $R^2 = 0.59$; song: $F(1, 585) = 126$, $R^2 = 0.18$; *both* $P < 0.0001$; P values computed using robust standard errors); and the grey ribbons represent the standard errors of the mean, from the regressions.

as infant-directed but these features did not reliably correlate with infant-directed song across cultures in the corpus (and, accordingly, neither approach to the acoustic analyses identified them as reliable correlates of infant-directedness in music). Intuitions concerning infant-directed song may also have been driven by more subjective features of the recordings, higher-level acoustic features that we did not measure or both.

We note, however, that the interpretation of this difference may be limited by the representativeness of the sample of recordings: the differences in the ability of the model to predict listeners' intuitions could alternatively be driven by differences in the true representativeness of one or more of the vocalization types.

Discussion

We provide convergent evidence for cross-cultural regularities in the acoustic design of infant-directed speech and song. Infant-directedness was robustly characterized by core sets of acoustic features, across the 21 societies studied, and these sets of features differed reliably across speech and song. Naive listeners were sensitive to the acoustical regularities, as they reliably identified infant-directed vocalizations as more infant-directed than adult-directed vocalizations, despite the fact that the vocalizations were of largely unfamiliar cultural, geographic and linguistic origin.

Thus, despite evident variability in language, music and infant care practices worldwide, when people speak or sing to fussy infants, they modify the acoustic features of their vocalizations in similar and mutually intelligible ways across cultures. This evidence supports the hypothesis that the forms of infant-directed vocalizations are shaped by their functions, in a fashion similar to the vocal signals of many non-human species.

These findings do not mean that infant-directed speech and song always sound the same across cultures. Indeed, the classification accuracy of a machine-learning model varied, with some fieldsites demonstrating larger acoustic differences between infant- and adult-directed vocalizations than other fieldsites. Similarly, the citizen-science participants' ratings of infant-directedness differed substantially in magnitude across fieldsites. But such variability also does not imply the absence of cross-cultural regularities. Instead, the variability supports an account of acoustic variation stemming from epigenetic rules: species-typical traits that bias cultural variation in one direction rather than another[86]. Put another way, the patterns of evidence reported here strongly imply a core set of cross-cultural acoustic and perceptual regularities that are also shaped by culture.

By analysing both speech and song recorded from the same voices, we discerned precise differences in the ways infant-directedness is instantiated in language and music. In response to the same prompt of addressing a 'fussy infant', infant-directedness in speech and song was instantiated with opposite trends in acoustic modification (relative to adult-directed speech and song, respectively): infant-directed speech was more intense and contrasting (for example, more pitch variability, higher intensity) while infant-directed song was more subdued and soothing (for example, less pitch variability, lower intensity). These acoustic dissociations comport with functional dissociations, with speech being more attention-grabbing, the better to distract from a baby's fussiness[37,38]; and song being more soothing, the better to lower a baby's arousal[32,33,41,77,78,83,84]. Speech and song are both capable of playful or soothing roles[53] but each here tended toward one acoustic profile over the other, despite both types of vocalization being elicited here in the same context: vocalizations used "when the baby is fussy".

Many of the reported acoustic differences are consistent with properties of vocal signalling in non-human animals, raising the intriguing possibility that the designs of human communication systems are rooted in the basic principles of bioacoustics[1–15]. For example, in both speech and song, infant-directedness was robustly associated with purer and less harsh vocal timbres and greater formant-frequency dispersion (expanded vowel space). And in speech, one of the largest and most cross-culturally robust effects of infant-directedness was higher pitch (F_0). In non-human animals, these features have convergently evolved across taxa in the

functional context of signalling friendliness or approachability in close contact calls[1,3,56,87], in contrast to alarm calls or signals of aggression, which are associated with low-pitched, rough sounds with less formant dispersal[4,88–90]. The use of these features in infant care may originate from signalling approachability to baby but may have later acquired further functions more specific to the human developmental context. For example, greater formant-frequency dispersion accentuates vowel contrasts, which could facilitate language acquisition[36,56,91–93]; and purer vocal timbre may facilitate communication by contrasting conspicuously with the acoustic context of infant cries[5] (for readers unfamiliar with infants, their cries are acoustically harsh[17,82]).

Such conspicuous contrasts may have the effect of altering speech to make it more song-like when interacting with infants, as Fernald[18] notes: "…the communicative force of [parental] vocalizations derive not from their arbitrary meanings in a linguistic code but more from their immediate musical power to arouse and alert, to calm, and to delight".

Comparisons of the acoustic effects across speech and song reported here support this idea. Infant-directedness altered the pitch level (F_0) of speech, bringing it roughly to a level typical of song, while also increasing pulse clarity. These characteristics of music have been argued to originate from elaborations to infant-directed vocalizations, where both use less harsh but more variable pitch patterns, more temporally variable and expansive vowel spaces and attention-orienting rhythmic cues to provide infants with ostensible 'flashy' signals of attention and prosocial friendliness[41,42,54,94,95]. Pitch alterations are not absent from infant-directed song, of course; in one study, mothers sang a song at higher pitch when producing a more playful rendition and a lower pitch when producing a more soothing rendition[76]. But on average, both infant- and adult-directed song, along with infant-directed speech, tend to be higher in pitch than adult-directed speech. In sum: the constellation of acoustic features that characterize infant-directedness in speech, across cultures, is rather musical.

The current study has several limitations, leaving open at least four sets of further questions. First, the results are suggestive of universality in the production of infant-directed vocalizations because the corpus covers a swath of geographic locations (21 societies on six continents), languages (12 language families) and different subsistence regimes (8 types) (Table 1). But the participants studied do not constitute a representative sample of humans, nor do the societies or languages studied constitute a representative sample of human societies or languages. Future work is needed to assess the validity of such a universality claim by studying infant-directed vocalizations in a wider range of human societies and by using phylogenetic methods to examine whether people in societies that are distantly related nonetheless produce similar infant-directed vocalizations.

Second, the naive listener experiment tested a large number of participants and covered a diverse set of countries and native languages, raising the possibility that results may generalize. But the results might not generalize, however, because the instructions of the experiment were presented in English, on an English-language website. Future work may determine their generality by testing perceived infant-directedness in multilingual experiments, to more accurately characterize cross-cultural variability in the perception of infant-directedness, and by testing listener intuitions among groups with reduced exposure to a given set of infant-directed vocalizations, such as very young infants or people from isolated, distantly related societies, as in related efforts[27,64,96]. Such research would benefit in particular from a focus on societies previously reported to have unusual vocalization practices, infant care practices or both[47,49–51] and would also clarify the extent to which convergent practices across cultures are due to cultural borrowing (in the many cases where societies are not fully isolated from the influence of global media).

Third, most prior studies of infant-directed vocalizations use elicited recordings[20,23,26,30,39,76], as did we. While this method may underestimate the differences between infant-directed and adult-directed vocalizations, whether and how elicited infant-directed speech and song differ from their naturalistic counterparts is poorly understood. Future work may explore this issue by analysing recordings of infant-directed vocalizations that are covertly and/or unobtrusively collected in a non-elicited manner, as in research using wearable recording devices for infants[73,97]. This may also resolve potential confounds caused by the wording of instructions to vocalizers.

Last, we note that speech and song are used in multiple contexts with infants, of which "addressing a fussy infant" is just one[18,34]. One curious finding may bear on general questions of the psychological functions of music: naive listeners displayed a bias toward 'adult' guesses for speech and 'baby' guesses for song, regardless of their actual targets. We speculate that listeners treated 'adult' and 'baby' as the default reference levels for speech and song, respectively, against which acoustic evidence was compared, a pattern consistent with theories that posit song as having a special connection to infant care in human psychology[33,42].

Methods

Vocalization corpus. We built a corpus of 1,615 recordings of infant-directed song, infant-directed speech, adult-directed song and adult-directed speech (all audio is available at https://doi.org/10.5281/zenodo.5525161). Participants ($n=410$) living in 21 societies (Fig. 1a and Table 1) produced each of these vocalizations, respectively, with a median of 15 participants per society (range 6–57). From those participants for whom information was available, most were female (86%) and nearly all were parents or grandparents of the focal infant (95%). Audio for one or more examples was unavailable from a small minority of participants, in cases of equipment failure or when the participant declined to complete the full recording session (25 recordings or 1.5% of the corpus were missing).

Recordings were collected by principal investigators and/or staff at their fieldsites, all using the same data collection protocol. They translated instructions to the native language of the participants, following the standard research practices at each site. There was no procedure for screening out participants but we encouraged our collaborators to collect data from parents rather than non-parents. Fieldsites were selected partly by convenience (via recruiting principal investigators at fieldsites with access to infants and caregivers) and partly to maximize cultural, linguistic and geographic diversity (Table 1).

For infant-directed song and infant-directed speech, participants were asked to sing and speak to their infant as if they were fussy, where 'fussy' could refer to anything from frowning or mild whimpering to a full tantrum. At no fieldsites were difficulties reported in the translation of the English word 'fussy', suggesting that participants understood it. For adult-directed speech, participants spoke to the researcher about a topic of their choice (for example, they described their daily routine). For adult-directed song, participants sang a song that was not intended for infants; they also stated what that song was intended for (for example, "a celebration song"). Participants vocalized in the primary language of their fieldsite, with a few exceptions (for example, when singing songs without words; or in locations that used multiple languages, such as Turku, which included both Finnish and Swedish speakers).

For most participants (90%) an infant was physically present during the recording (the infants were 48% female; age in months: *mean* = 11.40; s.d. = 7.61; range 0.5–48). When an infant was not present, participants were asked to imagine that they were vocalizing to their own infant or grandchild and simulated their infant-directed vocalizations (a brief discussion is in Supplementary Results).

In all cases, participants were free to determine the content of their vocalizations. This was intentional: imposing a specific content category on their vocalizations (for example, "sing a lullaby") would probably alter the acoustic features of their vocalizations, which are known to be influenced by experimental contexts[98]. Some participants produced adult-directed songs that shared features with the intended soothing nature of the infant-directed songs; data on the intended behavioural context of each adult-directed song are in Supplementary Table 4.

All recordings were made with Zoom H2n digital audio recorders, using foam windscreens (where available). To ensure that participants were audible along with researchers, who stated information about the participant and environment before and after the vocalizations, recordings were made with a 360° dual x–y microphone pattern. This produced two uncompressed stereo audio files (WAV) per participant at 44.1 kHz; we only analysed audio from the two-channel file on which the participant was loudest.

The principal investigator at each fieldsite provided standardized background data on the behaviour and cultural practices of the society (for example, whether there was access to mobile phones/television/radio and how commonly people

used infant-directed speech or song in their daily lives). Most items were based on variables included in the D-PLACE cross-cultural corpus[99].

The 21 societies varied widely in their characteristics, from cities with millions of residents (Beijing) to small-scale hunter-gatherer groups of as few as 35 people (Hadza). All of the small-scale societies studied had limited access to television, radio and the internet, mitigating against the influence of exposure to the music and/or infant care practices of other societies. Four of the small-scale societies (Nyangatom, Toposa, Sápara/Achuar and Mbendjele) were completely without access to these communication technologies.

The societies also varied in the prevalence of infant-directed speech and song in day-to-day life. The only site reported to lack infant-directed song in contemporary practice was the Quechuan/Aymaran site, although it was also noted that people from this site know infant-directed songs in Spanish and use other vocalizations to calm infants. Conversely, the Mbendjele BaYaka were noted to use infant-directed song but rarely used infant-directed speech. In most sites, the frequency of infant-directed song and speech varied. For example, among the Tsimane, song was reportedly infrequent in the context of infant care; when it appears, however, it is apparently used to soothe and encourage infants to sleep.

Our default strategy was to analyse all available audio from the corpus. In some cases, however, this was inadvisable (for example, in the naive listener experiment, when a listener might understand the language of the recording and make a judgement on the basis of the recording's linguistic content rather than its acoustic content); all exclusion decisions are explicitly stated throughout.

Acoustic analyses. *Acoustic feature extraction.* We manually extracted the longest continuous and uninterrupted section of audio from each recording (that is, isolating vocalizations by the participant from interruptions from other speakers, the infant and so on), using Adobe Audition. We then used the silence detection tool in Praat[100], with minimum sounding intervals at 0.1 s and minimum silent intervals at 0.3 s, to remove all portions of the audio where the participant was not speaking (that is, the silence between vocalization phrases). These were manually concatenated in Python, producing denoised recordings, which were subsequently checked manually to ensure minimal loss of content.

We extracted and subsequently analysed acoustic features using Praat[100] and MIRtoolbox[101] and computed additional rhythm features using discrete Fourier transforms of the signal[102] and normalized pairwise variability of syllabic events[103]. These features consisted of measurements of pitch (for example, F_0, the fundamental frequency), timbre (for example, roughness) and rhythm (for example, tempo; note, because temporal measures would be affected by the concatenation process, we computed these variables on unconcatenated audio only); all summarized over time: producing 94 variables in total. We standardized feature values within-voices, eliminating between-voice variability. Further technical details are in Supplementary Methods.

For both the LASSO analyses (Fig. 1b) and the regression-based acoustic analyses (Fig. 2), we restricted the variable set to 27 summary statistics of median and IQR, as these correlated highly with other summary statistics (for example, maximum, range) but were less sensitive to extreme observations.

The LASSO modelling, mixed-effect modelling and PCA analysis were all run on the full corpus with only a few exceptions: we excluded ten recordings due to missing values on one or more acoustic features and a further 35 recordings where one or more recording was missing from the same vocalist, leaving 1,570 recordings for the analysis.

LASSO modelling. We trained least absolute shrinkage and selection operator (LASSO) logistic classifiers with cross-validation using tidymodels[104]. For both speech and song, these models were provided with the set of 27 acoustic variables described in the previous section. These raw features were then demeaned for speech and song separately within-voices and then normalized at the level of the whole corpus. During model training, multinomial log-loss was used as an evaluation metric to fit the lambda parameter of the model.

For the main analyses (Fig. 1b, Supplementary Table 2 and Supplementary Fig. 2) we used a k-fold cross-validation procedure at the level of fieldsites. Alternate approaches used k-fold cross-validation at the levels of language family and world region (Supplementary Fig. 1). We evaluated model performance using a receiver operating characteristic metric, binary AUC. This metric is commonly used to evaluate the diagnostic ability of a binary classifier; it yields a score between 0% and 100%, with a chance level of 50%.

Mixed-effects modelling. Following a preregistered exploratory–confirmatory design, we fitted a multilevel mixed-effects regression predicting each acoustic variable from the vocalization types, after adjusting for voice and fieldsite as random effects and allowing them to vary for each vocalization type separately. To reduce the risk of Type I error, we performed this analysis on a randomly selected half of the corpus (exploratory, weighting by fieldsite) and only report results that successfully replicated in the other half (confirmatory). We did not correct for multiple tests because the exploratory–confirmatory design restricts the tests to those with a directional prediction.

These analyses deviated from the preregistration in two minor ways. First, we retained planned comparisons within vocalization types, but we eliminated those that compared across speech and song when we found much larger acoustic differences between speech and song overall than the differences between infant- and adult-directed vocalizations (a fact we failed to predict). As such, we adopted the simpler approach of post-hoc comparisons that were only within speech and within song. For transparency, we still report the preregistered post-hoc tests in Supplementary Fig. 4 but suggest that these comparisons be interpreted with caution. Second, to enable fieldsite-wise estimates (reported in Extended Data Fig. 1), we normalized the acoustic data corpus-wide and included a random effect of participant, rather than normalizing within-voices (as within-voice normalization would set all fieldsite-level effects to 0, making cross-fieldsite comparisons impossible).

Naive listener experiment. We analysed all data available at the time of writing this paper from the 'Who's Listening?' game at https://themusiclab.org/quizzes/ids, a continuously running jsPsych[105] experiment distributed via Pushkin[106], a platform that facilitates large-scale citizen-science research. This approach involves the recruitment of volunteer participants, who typically complete experiments because the experiments are intrinsically rewarding, with larger and more diverse samples than are typically feasible with in-laboratory research[85,107]. A total of 68,206 participants began the experiment, the first in January 2019 and the last in October 2021. Demographics in the subsample of United States participants are in Supplementary Table 6.

We played participants vocalizations from a subset of the corpus, excluding those that were less than 10 s in duration ($n=111$) and those with confounding sounds produced by a source other than the target voice in the first 5 s of the recording (for example, a crying baby or laughing adult in the background; $n=366$), as determined by two independent annotators who remained unaware of vocalization type and fieldsite with disagreements resolved by discussion. A test of the robustness of the main effects to this exclusion decision is in Supplementary Results. We also excluded participants who reported having previously participated in the same experiment ($n=3,889$), participants who reported being younger than 12 years old ($n=1,519$) and those who reported having a hearing impairment ($n=1,437$).

This yielded a sample of 51,065 participants (gender: 22,862 female, 27,045 male, 1,117 other, 41 did not disclose; age: median 22 yr, IQR 18–29). Participants self-reported living in 187 different countries (Fig. 3b) and self-reported speaking 172 first languages and 147 second languages (27 of which were not in the list of first languages), for a total of 199 different languages. Roughly half the participants were native English speakers from the United States. We supplemented these data with a paid online experiment, to increase the sampling of a subset of recordings in the corpus (Supplementary Methods).

Participants listened to at least 1 and at most 16 vocalizations drawn from the subset of the corpus (as they were free to leave the experiment before completing it) for a total of 495,512 ratings (infant-directed song: $n=139,708$; infant-directed speech: $n=99,482$; adult-directed song: $n=132,124$; adult-directed speech: $n=124,198$). The vocalizations were selected with blocked randomization, such that a set of 16 trials included 4 vocalizations in English and 12 in other languages; this method ensured that participants heard a substantial number of non-English vocalizations. This yielded a median of 516.5 ratings per vocalization (IQR 315–566; range 46–704) and thousands of ratings for each society (median = 22,974; IQR 17,458–25,177). The experiment was conducted only in English, so participants probably had at least a cursory knowledge of English; a test of the robustness of the main effects when excluding English-language recordings is in Supplementary Results.

We asked participants to classify each vocalization as directed toward either a baby or an adult. The prompt 'Someone is speaking or singing. Who do you think they are singing or speaking to?' was displayed while the audio played; participants could respond with either 'adult' or 'baby', by pressing a key corresponding to either a drawing of an infant or an adult face (when the participant used a desktop computer) or by tapping one of the faces (when the participant used a tablet or smartphone). The locations of the faces (left versus right on a desktop; top versus bottom on a tablet or smartphone) were randomized participant-wise. Screenshots are in Extended Data Fig. 3.

We asked participants to respond as quickly as possible, a common instruction in perception experiments, to reduce variability that could be introduced by participants hearing differing lengths of each stimulus, to reduce the likelihood that participants used linguistic content to inform their decisions and to facilitate a response-time analysis (Extended Data Fig. 5), as jsPsych provides reliable response-time data[108]. We also used the response-time data as a coarse measure of compliance, by dropping trials where participants were probably inattentive, responding very quickly (<500 ms) or slowly (>5 s). Most response times fell within this time window (82.1% of trials).

The experiment included two training trials, using English-language recordings of a typically infant-directed song ('The wheels on the bus') and a typically adult-directed song ('Hallelujah'); 92.7% of participants responded correctly by the first try and 99.5% responded correctly by the second try, implying that the vast majority of the participants understood the task.

As soon as they made a choice, playback stopped. After each trial, we told participants whether or not they had answered correctly and how long, in seconds,

they took to respond. At the end of the experiment, we showed participants their total score and percentile rank (relative to other participants).

Ethics. Ethics approval for the collection of recordings was provided by local institutions and/or the home institution of the collaborating author who collected data at each fieldsite. These included the Bioethics Committee, Jagiellonian University (1072.6120.48.2017); Board for Research Ethics, Åbo Akademi University; Committee on the Use of Human Subjects, Harvard University (IRB16-1080 and IRB18-1739); Ethics Committee, School of Psychology, Victoria University of Wellington (0000023076); Human Investigation Committee, Yale University (MODCR00000571); Human Participants Ethics Committee, University of Auckland (018981); Human Research Protections Program, University of California, San Diego (161173); Institutional Review Board, Arizona State University (STUDY00008158); Institutional Review Board, Florida International University (IRB17-0067); Institutional Review Board, Future Generations University; Max Planck Institute for Evolutionary Anthropology; Research Ethics Board, University of Toronto (33547); Research Ethics Committee, University College, London (13121/001); Review Board for Ethical Standards in Research, Toulouse School of Economics/IAST (2017-06-001 and 2018-09-001); and Tanzania Commission for Science and Technology (COSTECH). Ethics approval for the naive listener experiment was provided by the Committee on the Use of Human Subjects, Harvard University (IRB17-1206). Informed consent was obtained from all participants.

Statistics and reproducibility. All data and code are provided (see the Data availability and Code availability statements). The sample sizes were not chosen a priori for either the participants who provided recordings or the participants in the naive listener experiment. All data exclusions are fully reported (see the corresponding Methods sections, above) and these decisions were either made before the analyses being conducted (for example, excluding naive listeners reporting hearing impairment) or, for post hoc exclusion decisions, were justified by subsequent analyses (for example, when a confound was discovered after the fact). For an example of the latter, to compute d' scores at the level of each vocalist, both infant-directed and adult-directed versions of a vocalization (speech or song) were required, so we excluded the small number of vocalists that did not have complete pairs. The experiment did not involve any randomization of conditions or experimenter blinding, although the selection of recordings the participants heard was randomized. For all statistical tests, assumptions were assessed visually; when potential violations to normality of residuals were detected, we used robust standard errors to compute P values.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The audio corpus is available at https://doi.org/10.5281/zenodo.5525161. All data, including supplementary fieldsite-level data and the recording collection protocol, are available at https://github.com/themusiclab/infant-speech-song and are permanently archived at https://doi.org/10.5281/zenodo.6562398. The preregistration for the auditory analyses is at https://osf.io/5r72u.

Code availability

Analysis and visualization code, a reproducible R Markdown manuscript and code for the naive listener experiment are available at https://github.com/themusiclab/infant-speech-song and are permanently archived at https://doi.org/10.5281/zenodo.6562398.

Received: 3 May 2022; Accepted: 10 June 2022;
Published online: 18 July 2022

References

1. Morton, E. S. On the occurrence and significance of motivation–structural rules in some bird and mammal sounds. *Am. Nat.* **111**, 855–869 (1977).
2. Endler, J. A. Some general comments on the evolution and design of animal communication systems. *Phil. Trans. R. Soc. B* **340**, 215–225 (1993).
3. Owren, M. J. & Rendall, D. Sound on the rebound: bringing form and function back to the forefront in understanding nonhuman primate vocal signaling. *Evol. Anthropol.* **10**, 58–71 (2001).
4. Fitch, W. T., Neubauer, J. & Herzel, H. Calls out of chaos: the adaptive significance of nonlinear phenomena in mammalian vocal production. *Anim. Behav.* **63**, 407–418 (2002).
5. Wiley, R. H. The evolution of communication: information and manipulation. *Anim. Behav.* **2**, 156–189 (1983).
6. Krebs, J. & Dawkins, R. Animal signals: Mind-reading and manipulation. In *Behavioural Ecology: An Evolutionary Approach* (eds Krebs, J. & Davies, N.) 380–402 (Blackwell, 1984).
7. Karp, D., Manser, M. B., Wiley, E. M. & Townsend, S. W. Nonlinearities in meerkat alarm calls prevent receivers from habituating. *Ethology* **120**, 189–196 (2014).
8. Slaughter, E. I., Berlin, E. R., Bower, J. T. & Blumstein, D. T. A test of the nonlinearity hypothesis in great-tailed grackles (*Quiscalus mexicanus*). *Ethology* **119**, 309–315 (2013).
9. Wagner, W. E. Fighting, assessment, and frequency alteration in Blanchard's cricket frog. *Behav. Ecol. Sociobiol.* **25**, 429–436 (1989).
10. Ladich, F. Sound production by the river bullhead, *Cottus gobio* L. (Cottidae, Teleostei). *J. Fish Biol.* **35**, 531–538 (1989).
11. Filippi, P. et al. Humans recognize emotional arousal in vocalizations across all classes of terrestrial vertebrates: evidence for acoustic universals. *Proc. R. Soc. B.* **284**, 20170990 (2017).
12. Lingle, S. & Riede, T. Deer mothers are sensitive to infant distress vocalizations of diverse mammalian species. *Am. Nat.* **184**, 510–522 (2014).
13. Custance, D. & Mayer, J. Empathic-like responding by domestic dogs (*Canis familiaris*) to distress in humans: an exploratory study. *Anim. Cogn.* **15**, 851–859 (2012).
14. Lea, A. J., Barrera, J. P., Tom, L. M. & Blumstein, D. T. Heterospecific eavesdropping in a nonsocial species. *Behav. Ecol.* **19**, 1041–1046 (2008).
15. Magrath, R. D., Haff, T. M., McLachlan, J. R. & Igic, B. Wild birds learn to eavesdrop on heterospecific alarm calls. *Curr. Biol.* **25**, 2047–2050 (2015).
16. Piantadosi, S. T. & Kidd, C. Extraordinary intelligence and the care of infants. *Proc. Natl Acad. Sci. USA* **113**, 6874–6879 (2016).
17. Soltis, J. The signal functions of early infant crying. *Behav. Brain Sci.* **27**, 443–458 (2004).
18. Fernald, A. Human maternal vocalizations to infants as biologically relevant signals: An evolutionary perspective. In *The Adapted Mind: Evolutionary Psychology and the Generation of Culture* (eds Barkow, J. H. et al.) 391–428 (Oxford Univ. Press, 1992).
19. Burnham, E., Gamache, J. L., Bergeson, T. & Dilley, L. Voice-onset time in infant-directed speech over the first year and a half. *Proc. Mtgs Acoust.* **19**, 060094 (2013).
20. Fernald, A. & Mazzie, C. Prosody and focus in speech to infants and adults. *Dev. Psychol.* **27**, 209–221 (1991).
21. Ferguson, C. A. Baby talk in six languages. *Am. Anthropol.* **66**, 103–114 (1964).
22. Audibert, N. & Falk, S. Vowel space and f0 characteristics of infant-directed singing and speech. In *Proc. 9th International Conference on Speech Prosody.* 153–157 (2018).
23. Kuhl, P. K. et al. Cross-language analysis of phonetic units in language addressed to infants. *Science* **277**, 684–686 (1997).
24. Englund, K. T. & Behne, D. M. Infant directed speech in natural interaction: Norwegian vowel quantity and quality. *J. Psycholinguist. Res.* **34**, 259–280 (2005).
25. Fernald, A. The perceptual and affective salience of mothers' speech to infants. In *The Origins and Growth of Communication* (eds Feagans, L. et al.) 5–29 (Praeger, 1984).
26. Falk, S. & Kello, C. T. Hierarchical organization in the temporal structure of infant-direct speech and song. *Cognition* **163**, 80–86 (2017).
27. Bryant, G. A. & Barrett, H. C. Recognizing intentions in infant-directed speech: evidence for universals. *Psychol. Sci.* **18**, 746–751 (2007).
28. Piazza, E. A., Iordan, M. C. & Lew-Williams, C. Mothers consistently alter their unique vocal fingerprints when communicating with infants. *Curr. Biol.* **27**, 3162–3167 (2017).
29. Trehub, S. E., Unyk, A. M. & Trainor, L. J. Adults identify infant-directed music across cultures. *Infant Behav. Dev.* **16**, 193–211 (1993).
30. Trehub, S. E., Unyk, A. M. & Trainor, L. J. Maternal singing in cross-cultural perspective. *Infant Behav. Dev.* **16**, 285–295 (1993).
31. Mehr, S. A., Singh, M., York, H., Glowacki, L. & Krasnow, M. M. Form and function in human song. *Curr. Biol.* **28**, 356–368 (2018).
32. Mehr, S. A. et al. Universality and diversity in human song. *Science* **366**, 957–970 (2019).
33. Trehub, S. E. Musical predispositions in infancy. *Ann. NY Acad. Sci.* **930**, 1–16 (2001).
34. Trehub, S. E. & Trainor, L. Singing to infants: lullabies and play songs. *Adv. Infancy Res.* **12**, 43–78 (1998).
35. Trehub, S. E. et al. Mothers' and fathers' singing to infants. *Dev. Psychol.* **33**, 500–507 (1997).
36. Thiessen, E. D., Hill, E. A. & Saffran, J. R. Infant-directed speech facilitates word segmentation. *Infancy* **7**, 53–71 (2005).
37. Trainor, L. J. & Desjardins, R. N. Pitch characteristics of infant-directed speech affect infants' ability to discriminate vowels. *Psychon. Bull. Rev.* **9**, 335–340 (2002).
38. Werker, J. F. & McLeod, P. J. Infant preference for both male and female infant-directed talk: a developmental study of attentional and affective responsiveness. *Can. J. Psychol.* **43**, 230–246 (1989).
39. Ma, W., Fiveash, A., Margulis, E. H., Behrend, D. & Thompson, W. F. Song and infant-directed speech facilitate word learning. *Q. J. Exp. Psychol.* **73**, 1036–1054 (2020).
40. Falk, D. Prelinguistic evolution in early hominins: whence motherese? *Behav. Brain Sci.* **27**, 491–502 (2004).

41. Mehr, S. A. & Krasnow, M. M. Parent–offspring conflict and the evolution of infant-directed song. *Evol. Hum. Behav.* **38**, 674–684 (2017).
42. Mehr, S. A., Krasnow, M. M., Bryant, G. A. & Hagen, E. H. Origins of music in credible signaling. *Behav. Brain Sci.* https://doi.org/10.1017/S0140525X20000345 (2020).
43. Senju, A. & Csibra, G. Gaze following in human infants depends on communicative signals. *Curr. Biol.* **18**, 668–671 (2008).
44. Hernik, M. & Broesch, T. Infant gaze following depends on communicative signals: an eye-tracking study of 5- to 7-month-olds in Vanuatu. *Dev. Sci.* **22**, e12779 (2019).
45. Henrich, J., Heine, S. J. & Norenzayan, A. The weirdest people in the world? *Behav. Brain Sci.* **33**, 61–83 (2010).
46. Yarkoni, T. The generalizability crisis. *Behav. Brain Sci.* **45**, e1 (2022).
47. Broesch, T. & Bryant, G. A. Fathers' infant-directed speech in a small-scale society. *Child Dev.* **89**, e29–e41 (2018).
48. Ochs, E. & Schieffelin, B. Language acquisition and socialization. In *Culture Theory: Essays on Mind, Self, and Emotion* (eds Shweder, R. A. & LeVine, R. A.) 276–320 (Cambridge Univ. Press, 1984).
49. Schieffelin, B. B. *The Give and Take of Everyday Life: Language, Socialization of Kaluli Children* (Cambridge Univ. Press Archive, 1990).
50. Ratner, N. B. & Pye, C. Higher pitch in BT is not universal: acoustic evidence from Quiche Mayan. *J. Child Lang.* **11**, 515–522 (1984).
51. Pye, C. Quiché Mayan speech to children. *J. Child Lang.* **13**, 85–100 (1986).
52. Heath, S. B. *Ways with Words: Language, Life and Work in Communities and Classrooms* (Cambridge Univ. Press, 1983).
53. Trehub, S. E. Challenging infant-directed singing as a credible signal of maternal attention. *Behav. Brain Sci.* **44**, e117 (2021).
54. Räsänen, O., Kakouros, S. & Soderstrom, M. Is infant-directed speech interesting because it is surprising? Linking properties of IDS to statistical learning and attention at the prosodic level. *Cognition* **178**, 193–206 (2018).
55. Cristia, A. & Seidl, A. The hyperarticulation hypothesis of infant-directed speech. *J. Child Lang.* **41**, 913–934 (2014).
56. Kalashnikova, M., Carignan, C. & Burnham, D. The origins of babytalk: smiling, teaching or social convergence? *R. Soc. Open Sci.* **4**, 170306 (2017).
57. Grieser, D. L. & Kuhl, P. K. Maternal speech to infants in a tonal language: support for universal prosodic features in motherese. *Dev. Psychol.* **24**, 14 (1988).
58. Fisher, C. & Tokura, H. Acoustic cues to grammatical structure in infant-directed speech: cross-linguistic evidence. *Child Dev.* **67**, 3192–3218 (1996).
59. Kitamura, C., Thanavishuth, C., Burnham, D. & Luksaneeyanawin, S. Universality and specificity in infant-directed speech: pitch modifications as a function of infant age and sex in a tonal and non-tonal language. *Infant Behav. Dev.* **24**, 372–392 (2001).
60. Fernald, A. Intonation and communicative intent in mothers' speech to infants: is the melody the message? *Child Dev.* **60**, 1497–1510 (1989).
61. Farran, L. K., Lee, C.-C., Yoo, H. & Oller, D. K. Cross-cultural register differences in infant-directed speech: an initial study. *PLoS ONE* **11**, e0151518 (2016).
62. Broesch, T. L. & Bryant, G. A. Prosody in infant-directed speech is similar across Western and traditional cultures. *J. Cogn. Dev.* **16**, 31–43 (2015).
63. Broesch, T., Rochat, P., Olah, K., Broesch, J. & Henrich, J. Similarities and differences in maternal responsiveness in three societies: evidence From Fiji, Kenya, and the United States. *Child Dev.* **87**, 700–711 (2016).
64. ManyBabies Consortium. Quantifying sources of variability in infancy research using the infant-directed-speech preference. *Adv. Methods Pract. Psychol. Sci.* **3**, 24–52 (2020).
65. Soley, G. & Sebastian-Galles, N. Infants' expectations about the recipients of infant-directed and adult-directed speech. *Cognition* **198**, 104214 (2020).
66. Byers-Heinlein, K. et al. A multilab study of bilingual infants: exploring the preference for infant-directed speech. *Adv. Methods. Pract. Psychol. Sci.* https://doi.org/10.1177/2515245920974622 (2021).
67. Fernald, A. et al. A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *J. Child Lang.* **16**, 477–501 (1989).
68. Kitamura, C. & Burnham, D. Pitch and communicative intent in mother's speech: adjustments for age and sex in the first year. *Infancy* **4**, 85–110 (2003).
69. Kitamura, C. & Lam, C. Age-specific preferences for infant-directed affective intent. *Infancy* **14**, 77–100 (2009).
70. Hilton, C., Crowley, L., Yan, R., Martin, A. & Mehr, S. Children infer the behavioral contexts of unfamiliar foreign songs. Preprint at *PsyArXiv* https://doi.org/10.31234/osf.io/rz6qn (2021).
71. Yan, R. et al. Across demographics and recent history, most parents sing to their infants and toddlers daily. *Phil. Trans. R. Soc. B.* **376** (2021).
72. Custodero, L. A., Rebello Britto, P. & Brooks-Gunn, J. Musical lives: a collective portrait of American parents and their young children. *J. Appl. Dev. Psychol.* **24**, 553–572 (2003).
73. Mendoza, J. K. & Fausey, C. M. Everyday music in infancy. *Developmental Science*, **24** (2021).
74. Konner, M. Aspects of the developmental ethology of a foraging people. In *Ethological Studies of Child Behaviour* (ed. Blurton Jones, N. G.) 285–304 (Cambridge Univ. Press, 1972).
75. Marlowe, F.*The Hadza Hunter-Gatherers of Tanzania* (Univ. of California Press, 2010).
76. Cirelli, L. K., Jurewicz, Z. B. & Trehub, S. E. Effects of maternal singing style on mother–infant arousal and behavior. *J. Cogn. Neurosci.* **32**, 1213–1220 (2020).
77. Cirelli, L. K. & Trehub, S. E. Familiar songs reduce infant distress. *Dev. Psychol.* **56**, 861–868 (2020). https://doi.org/10.1037/dev0000917
78. Bainbridge, C. M. et al. Infants relax in response to unfamiliar foreign lullabies. *Nat. Hum. Behav.* **5**, 256–264 (2021).
79. Friedman, J., Hastie, T. & Tibshirani, R. Lasso and elastic-net regularized generalized linear models. R package version 2.0-5 (2016).
80. Hagen, E. H. & Bryant, G. A. Music and dance as a coalition signaling system. *Hum. Nat.* **14**, 21–51 (2003).
81. Corbeil, M., Trehub, S. E. & Peretz, I. Singing delays the onset of infant distress. *Infancy* **21**, 373–391 (2016).
82. Arnal, L. H., Flinker, A., Kleinschmidt, A., Giraud, A.-L. & Poeppel, D. Human screams occupy a privileged niche in the communication soundscape. *Curr. Biol.* **25**, 2051–2056 (2015).
83. Mehr, S. A., Kotler, J., Howard, R. M., Haig, D. & Krasnow, M. M. Genomic imprinting is implicated in the psychology of music. *Psychol. Sci.* **28**, 1455–1467 (2017).
84. Kotler, J., Mehr, S. A., Egner, A., Haig, D. & Krasnow, M. M. Response to vocal music in Angelman syndrome contrasts with Prader–Willi syndrome. *Evol. Hum. Behav.* **40**, 420–426 (2019).
85. Hilton, C. B. & Mehr, S. A. Citizen science can help to alleviate the generalizability crisis. *Behav. Brain Sci.* **45**, e21 (2022).
86. Lumsden, C. J. & Wilson, E. O. Translation of epigenetic rules of individual behavior into ethnographic patterns. *Proc. Natl Acad. Sci. USA* **77**, 4382–4386 (1980).
87. Fitch, W. T. Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. *J. Acoust. Soc. Am.* **102**, 1213 (1997).
88. Blumstein, D. T., Bryant, G. A. & Kaye, P. The sound of arousal in music is context-dependent. *Biol. Lett.* **8**, 744–747 (2012).
89. Reber, S. A. et al. Formants provide honest acoustic cues to body size in American alligators. *Sci. Rep.* **7**, 1816 (2017).
90. Reby, D. et al. Red deer stags use formants as assessment cues during intrasexual agonistic interactions. *Proc. R. Soc. B* **272**, 941–947 (2005).
91. Bertoncini, J., Jusczyk, P. W., Kennedy, L. J. & Mehler, J. An investigation of young infants' perceptual representations of speech sounds. *J. Exp. Psychol. Gen.* **117**, 21–33 (1988).
92. Werker, J. F. & Lalonde, C. E. Cross-language speech perception: initial capabilities and developmental change. *Dev. Psychol.* **24**, 672 (1988).
93. Polka, L. & Werker, J. F. Developmental changes in perception of nonnative vowel contrasts. *J. Exp. Psychol. Hum. Percept. Perform.* **20**, 421–435 (1994).
94. Trainor, L. J., Clark, E. D., Huntley, A. & Adams, B. A. The acoustic basis of preferences for infant-directed singing. *Infant Behav. Dev.* **20**, 383–396 (1997).
95. Tsang, C. D., Falk, S. & Hessel, A. Infants prefer infant-directed song over speech. *Child Dev.* **88**, 1207–1215 (2017).
96. McDermott, J. H., Schultz, A. F., Undurraga, E. A. & Godoy, R. A. Indifference to dissonance in native Amazonians reveals cultural variation in music perception. *Nature* **535**, 547–550 (2016).
97. Bergelson, E. et al. Everyday language input and production in 1001 children from 6 continents. Preprint at *PsyArXiv* https://doi.org/10.31234/osf.io/fjr5q (2022).
98. Trehub, S. E., Hill, D. S. & Kamenetsky, S. B. Parents' sung performances for infants. *Can. J. Exp. Psychol.* **51**, 385–396 (1997).
99. Kirby, K. R. et al. D-PLACE: a global database of cultural, linguistic and environmental diversity. *PLoS ONE* **11**, e0158391 (2016).
100. Boersma, P. Praat, a system for doing phonetics by computer. *Glot. Int.* **5**, 341–345 (2001).
101. Lartillot, O., Toiviainen, P. & Eerola, T. A Matlab toolbox for music information retrieval. In *Data Analysis, Machine Learning and Applications* (eds Preisach, C. et al.) 261–268 (Springer, 2008).
102. Patel, A. D. Musical rhythm, linguistic rhythm, and human evolution. *Music Percept.* **24**, 99–104 (2006).
103. Mertens, P. The prosogram: semi-automatic transcription of prosody based on a tonal perception model. In *Proc. 2nd International Conference on Speech Prosody* (eds Bel, B. & Marlien, I.) 549–552 (ISCA, 2004).
104. Kuhn, M. & Wickham, H. Tidymodels: A collection of packages for modeling and machine learning using tidyverse principles. R package version 0.2.0 (2020).
105. de Leeuw, J. R. jsPsych: a JavaScript library for creating behavioral experiments in a Web browser. *Behav. Res. Methods* **47**, 1–12 (2015).
106. Hartshorne, J. K., de Leeuw, J., Goodman, N., Jennings, M. & O'Donnell, T. J. A thousand studies for the price of one: accelerating psychological science with Pushkin. *Behav. Res. Methods* **51**, 1782–1803 (2019).

107. Sheskin, M. et al. Online developmental science to foster innovation, access, and impact. *Sci. Soc.* **24**, 675-678 (2020).
108. Anwyl-Irvine, A., Dalmaijer, E. S., Hodges, N. & Evershed, J. K. Realistic precision and accuracy of online experiment platforms, web browsers, and devices. *Behav. Res. Methods* **53**, 1407–1425 (2021).

Acknowledgements

This research was supported by the Harvard University Department of Psychology (M.M.K. and S.A.M.); the Harvard College Research Program (H.L.-R.); the Harvard Data Science Initiative (S.A.M.); the National Institutes of Health Director's Early Independence Award DP5OD024566 (S.A.M. and C.B.H.); the Academy of Finland grant no. 298513 (J. Antfolk); the Royal Society of New Zealand Te Aparangi Rutherford Discovery Fellowship RDF-UOA1101 (Q.D.A. and T.A.V.); the Social Sciences and Humanities Research Council of Canada (L.K.C.); the Polish Ministry of Science and Higher Education grant no. N43/DBS/000068 (G.J.); the Fogarty International Center (P.M., A. Siddaiah and C.D.P.); the National Heart, Lung, and Blood Institute and the National Institute of Neurological Disorders and Stroke award no. D43 TW010540 (P.M. and A. Siddaiah); the National Institute of Allergy and Infectious Diseases award no. R15-AI128714-01 (P.M.); the Max Planck Institute for Evolutionary Anthropology (C.T.R. and C.M.); a British Academy Research Fellowship and grant no. SRG-171409 (G.D.S.); the Institute for Advanced Study in Toulouse, under an Agence nationale de la recherche grant, Investissements d'Avenir ANR-17-EURE-0010 (L.G. and J. Stieglitz); the Fondation Pierre Mercier pour la Science (C.S.); and the Natural Sciences and Engineering Research Council of Canada (S.E.T.). We thank the participants and their families for providing recordings; L. Sugiyama for supporting pilot data collection; J. Du, E. Pillsworth, P. Wiessner and J. Ziker for collecting or attempting to collect additional recordings; N. Nicolas for research assistance in the Republic of the Congo; Z. Jurewicz for research assistance in Toronto; M. Delfi and R. Sakaliou for research assistance in Indonesia; W. Naiou and A. Altrin for research assistance in Vanuatu; S. Atwood, A. Bergson, D. Li, L. Lopez and E. Radytė for project-wide research assistance; and J. Kominsky, L. Powell and L. Yurdum for feedback on the manuscript. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

S.A.M. and M.M.K. conceived of the research, provided funding and coordinated the recruitment of collaborators and creation of the corpus. S.A.M. and M.M.K. designed the protocol for collecting vocalization recordings with input from D.A., who piloted it in the field. L.G., A.G., G.J., C.T.R., M.B.N., A. Martin, L.K.C., S.E.T., J. Song, M.K., A. Siddaiah, T.A.V., Q.D.A., J. Antfolk, P.M., A. Schachner, C.D.P., G.D.S., S.K., M.S., S.A.C., J.Q.P., C.S., J. Stieglitz, C.M., R.R.S. and B.M.W collected the field recordings, with support from E.A., A. Salenius, J. Andelin, S.C.C., M.A. and A. Mabulla. S.A.M., C.M.B. and J. Simson designed and implemented the online experiment. C.J.M. and H.L.-R. processed all recordings and designed the acoustic feature extraction with S.A.M. and M.M.K.; C.M.B. provided associated research assistance. C.M. designed the fieldsite questionnaire with assistance from M.B. and C.J.M., who collected the data from the principal investigators. C.B.H. and S.A.M. led analyses, with additional contributions from C.J.M., M.B., D.K. and M.M.K. C.B.H. and S.A.M. designed the figures. C.B.H. wrote computer code, with contributions from S.A.M., C.J.M. and M.B. D.K. conducted code review. C.J.M., H.L.-R., M.M.K. and S.A.M. wrote the initial manuscript. C.B.H. and S.A.M. wrote the first revision, with contributions from C.J.M. and M.B. S.A.M. wrote the second and third revisions, with contributions from C.B.H. and C.J.M.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41562-022-01410-x.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-022-01410-x.

Correspondence and requests for materials should be addressed to Courtney B. Hilton, Cody J. Moser or Samuel A. Mehr.

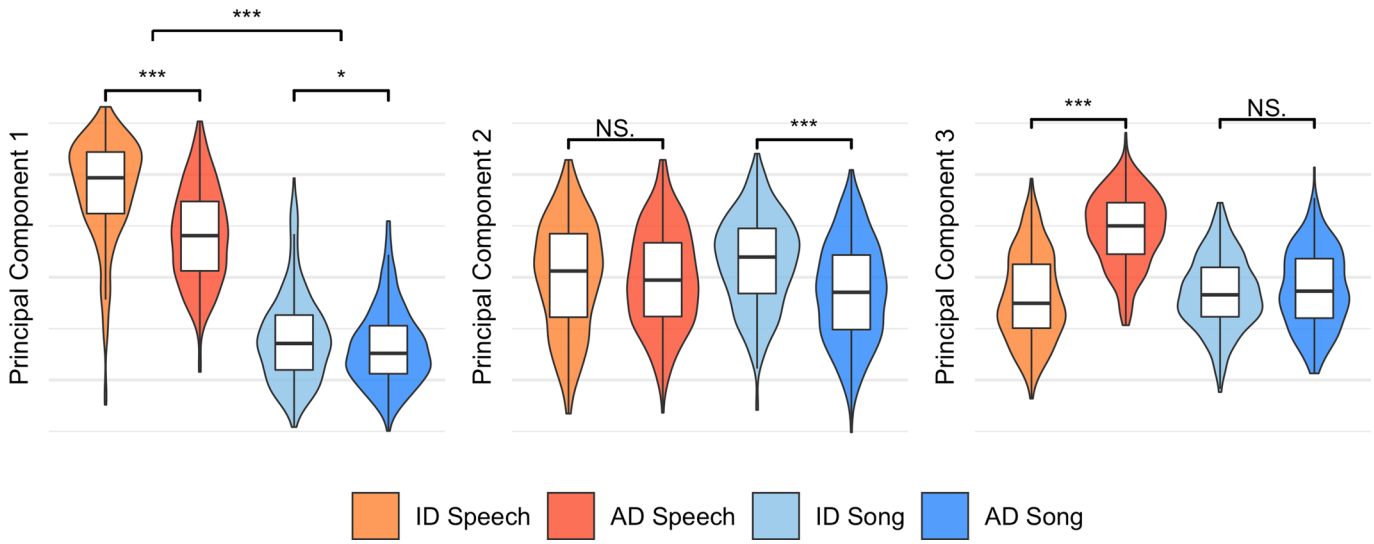
Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

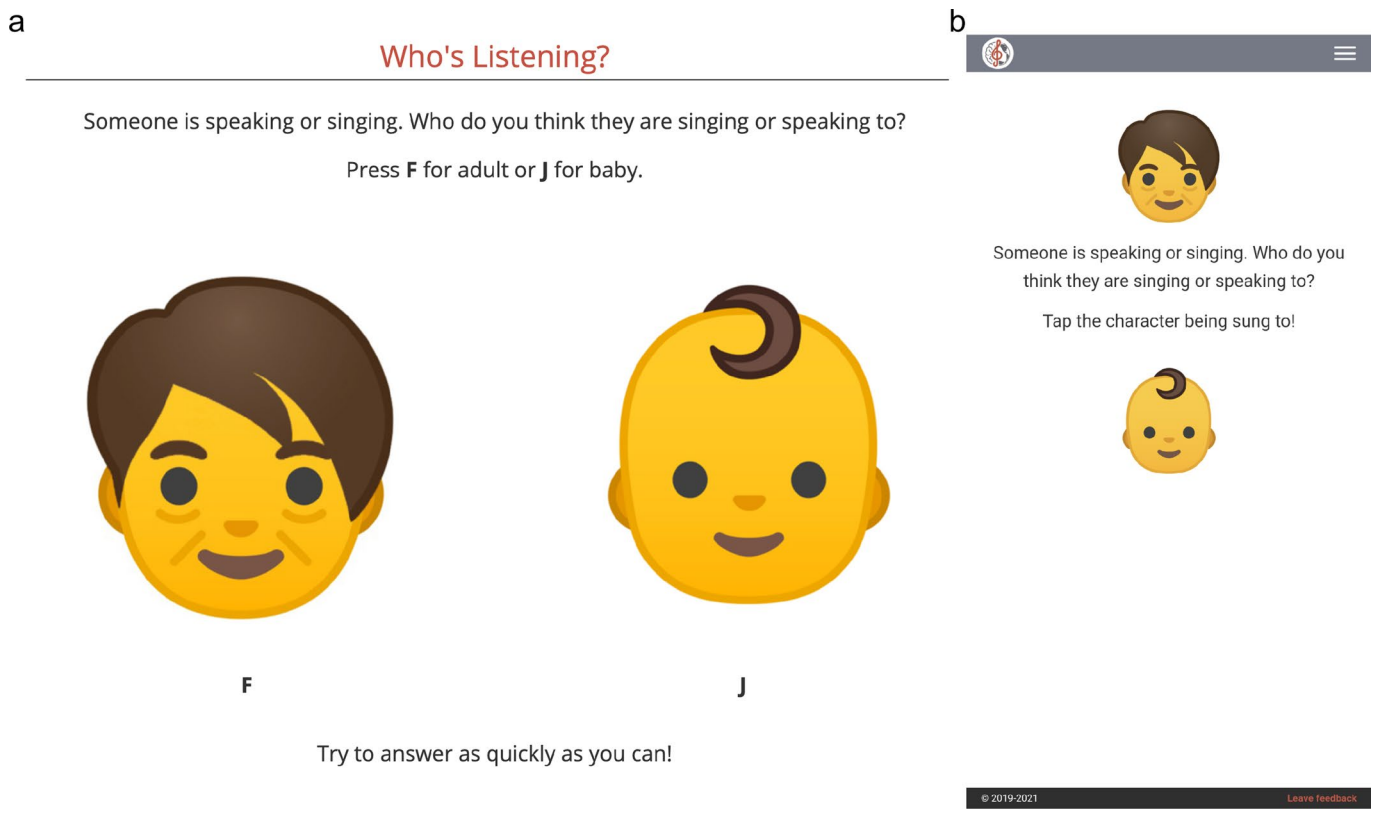
[1]Department of Psychology, Harvard University, Cambridge, MA, USA. [2]Haskins Laboratories, Yale University, New Haven, CT, USA. [3]Department of Cognitive and Information Sciences, University of California, Merced, Merced, CA, USA. [4]Boston College Department of Psychology, Chestnut Hill, MA, USA. [5]Department of Communication, University of California, Los Angeles, Los Angeles, CA, USA. [6]Department of Psychology, University of Amsterdam, Amsterdam, the Netherlands. [7]Operations, Information, and Decisions Department, The Wharton School of the University of Pennsylvania, Philadelphia, PA, USA. [8]Department of Anthropology, Boston University, Boston, MA, USA. [9]Jinka University, Jinka, Ethiopia. [10]Department of Environmental Health, Faculty of Health Sciences, Jagiellonian University Medical College, Krakow, Poland. [11]Department of Human Behavior, Ecology and Culture, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany. [12]School of Psychology, Victoria University of Wellington, Wellington, New Zealand. [13]Department of Philosophy, Classics, History of Art and Ideas, University of Oslo, Oslo, Norway. [14]Department of Psychology, University of Toronto, Scarborough, Toronto, Ontario, Canada. [15]Department of Psychology, University of Toronto, Mississauga, Mississauga, Ontario, Canada. [16]Department of Mathematics, University of California, Los Angeles, Los Angeles, CA, USA. [17]Department of Psychology, University of California, San Diego, La Jolla, CA, USA. [18]School of Psychology, University of Auckland, Auckland, New Zealand. [19]Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany. [20]Department of Psychology, Åbo Akademi, Turku, Finland. [21]Department of Health Promotion Sciences, College of Public Health, University of Arizona, Tucson, AZ, USA. [22]Department of Medicine, Division of Infectious Diseases, College of Medicine, University of Arizona, Tucson, AZ, USA. [23]Department of Family & Community Medicine, College of Medicine, University of Arizona, Tucson, AZ, USA. [24]Public Health Research Institute of India, Mysuru, India. [25]Department of Anthropology, Ball State University, Muncie, IN, USA. [26]Department of Anthropology, University College, London, London, UK. [27]Amsterdam Reproduction & Development, Amsterdam UMC, University of Amsterdam, Amsterdam, the Netherlands. [28]Department of Human Evolutionary Biology, Harvard University, Cambridge, MA, USA. [29]Institute for Advanced Study in Toulouse, Toulouse, France. [30]School of Human Evolution and Social Change, Arizona State University, Tempe, AZ, USA. [31]Division of Anthropology, California State University, Fullerton, CA, USA. [32]Institute of Evolutionary Medicine, University of Zurich, Zurich, Switzerland. [33]Université Toulouse, Toulouse, France. [34]Universidad Nacional del Altiplano Puno, Puno, Peru. [35]Department of Anthropology, University of California, Davis, Davis, CA, USA. [36]Centre for Culture & Evolution, Brunel University, London, Uxbridge, UK. [37]Future Generations University, Circle Ville, WV, USA. [38]Harpy Eagle Music Foundation, Georgetown, Guyana. [39]Mang'ola, Karatu, Tanzania. [40]Department of Archaeology and Heritage, University of Dar es Salaam, Dar es Salaam, Tanzania. [41]Department of Anthropology, University of California, Los Angeles, Los Angeles, CA, USA. [42]Division of Continuing Education, Harvard University, Cambridge, MA, USA. [43]Data Science Initiative, Harvard University, Cambridge, MA, USA. [44]These authors contributed equally: Courtney B. Hilton, Cody J. Moser. ✉e-mail: courtneyhilton@g.harvard.edu; cmoser2@ucmerced.edu; samuel.mehr@yale.edu

Acoustic Features	Afrocolombians	Arawak	Beijing	Hadza	Jenu Kurubas	Krakow	Rural Polish	Mbendjele	Meratwai Islanders	Colombian Mestizos	Nyangatom	Enga	Quechua	Sapara & Achuar	Topoisa	Toronto	Tsimane	Turku	San Diego	Tannese Vanuatuans	Wellington
Speech																					
Pulse Clarity	0.3	0.3	0.29	0.18	0.18	0.12	0.09	0.3	0.18	0.35	0.12	0.27	0.2	0.25	0.25	0.17	0.12	0.13	0.21	0.3	0.17
Energy Roll-Off (85th %-ile)	-0.58	-0.03	-0.32	-0.23	-0.24	-0.33	-0.45	-0.18	-0.19	-0.29	-0.13	-0.17	-0.22	-0.39	-0.09	-0.23	-0.33	-0.23	-0.41	-0.22	-0.24
Pitch (Median)	0.32	0.42	0.67	0.47	1.03	0.65	0.96	0.65	0.14	0.59	0.64	0.83	0.14	0.09	0.63	1.43	0.06	0.77	1.14	0.49	1.34
Inharmonicity	-0.31	-0.14	-0.34	-0.37	0	-0.29	-0.33	-0.18	-0.28	-0.41	-0.37	-0.15	-0.27	-0.14	-0.32	-0.43	0.02	-0.38	-0.26	-0.33	-0.46
Pitch (IQR)	0.22	0.67	0.53	0.22	0.78	0.87	1.11	0.33	0.06	0.52	0.23	0.67	0.26	0.21	0.17	1.82	-0.07	0.8	1.45	0.33	1.47
Vowel Travel Rate (Median)	0.44	0.15	0.07	0.84	0.89	0.66	0.89	-0.01	0.36	0.04	0.55	-0.31	0.33	0.58	0.1	1.4	0.36	0.77	1.12	0.35	1.2
Vowel Travel Rate (IQR)	0.22	0.11	0.14	0.91	1.04	0.77	0.89	0	0.2	-0.14	0.6	-0.15	0.24	0.5	0.12	1.39	0.4	0.88	1.15	0.39	1.24
Vowel Travel (IQR)	0.04	0.17	-0.2	0.33	0.37	0.43	0.74	-0.59	0.29	0.07	0.44	-0.03	-0.64	0.82	-0.42	1.01	0.26	0.8	0.89	0.32	0.95
Intensity (Median)	0.21	-0.06	0.11	0.27	0.4	-0.33	-0.09	0.27	0.11	0.13	0.13	0.15	0.2	0.25	0.13	-0.02	0.05	-0.15	-0.01	0.19	-0.27
Roughness (Median)	0.06	-0.56	-0.19	0.09	0.14	-1.19	-1.07	0.25	-0.06	0.09	-0.12	0.02	0.16	-0.01	-0.16	-0.28	0.04	-0.63	-0.57	-0.03	-0.61
Roughness (IQR)	0.1	-0.29	-0.17	0.11	0.17	-0.84	-0.46	0.15	-0.16	0.13	-0.23	0.02	0.09	0.15	-0.17	-0.12	-0.03	-0.35	-0.29	0.07	-0.5
Song																					
Intensity (Median)	-0.02	-0.23	-0.15	-0.01	0.1	-0.42	-0.31	-0.04	-0.09	-0.01	-0.21	-0.13	-0.02	0.11	-0.22	-0.1	-0.17	-0.22	-0.24	-0.04	-0.48
Vowel Travel (IQR)	0.19	0.29	0.18	-0.04	0.08	0.35	0.42	0.2	-0.06	0.22	0.08	0.21	-0.17	0.59	0.22	0.68	0.28	0.42	0.65	0.03	0.58
Vowel Travel Rate (IQR)	0.31	0.12	0.08	0.15	-0.03	0.37	0.43	1.18	-0.07	0.03	-0.18	0.41	0.15	0.16	0.17	0.34	0.14	0.14	0.24	-0.05	0.36
Roughness (IQR)	-0.08	-0.17	0.05	0	0.12	-1.07	-0.74	-0.03	-0.3	-0.03	-0.13	-0.18	0.08	0.29	-0.21	-0.21	0.08	-0.38	-0.41	-0.05	-0.6
Pitch (IQR)	-0.3	-0.14	-0.15	-0.29	0.02	-0.08	0.02	-0.4	-0.42	-0.14	-0.53	-0.14	-0.33	-0.36	-0.63	0.34	-0.41	-0.15	0.25	-0.31	0.12
Inharmonicity	-0.11	-0.16	-0.34	0.01	-0.31	-0.31	-0.22	-0.36	-0.23	0.25	0.16	-0.32	0.08	-0.38	0.15	-0.1	-0.63	-0.29	-0.44	-0.1	0.09
Vowel Travel Rate (Median)	0.08	-0.04	0.16	0.2	0	0.35	0.37	1.28	-0.08	-0.15	-0.32	0.55	0.08	0.07	0.09	0.37	0.16	0.11	0.22	-0.02	0.31
Roughness (Median)	-0.06	-0.39	0.04	-0.05	0.08	-1.16	-1.15	0.04	-0.17	-0.04	0.02	-0.13	0.17	0.17	-0.17	-0.36	0.15	-0.54	-0.56	-0.09	-0.56
Pulse Clarity	0.21	0.05	0.39	-0.14	0.09	-0.32	-0.35	0.34	-0.27	0.64	-0.23	0.44	-0.01	0.26	0.27	-0.02	-0.32	-0.37	0.1	0.48	0.12
Pitch (Median)	-0.29	0.03	-0.05	-0.01	0.06	0.08	0.15	-0.14	-0.11	-0.14	-0.36	-0.17	-0.21	0.2	-0.26	0.2	-0.09	0.02	0.22	-0.24	0.01
Energy Roll-Off (85th %-ile)	-0.49	0.22	-0.14	0.13	0.01	-0.16	-0.33	0.09	0.1	-0.11	0.18	0.18	-0.08	-0.19	0.26	0.03	-0.02	0.03	-0.22	0.04	-0.02

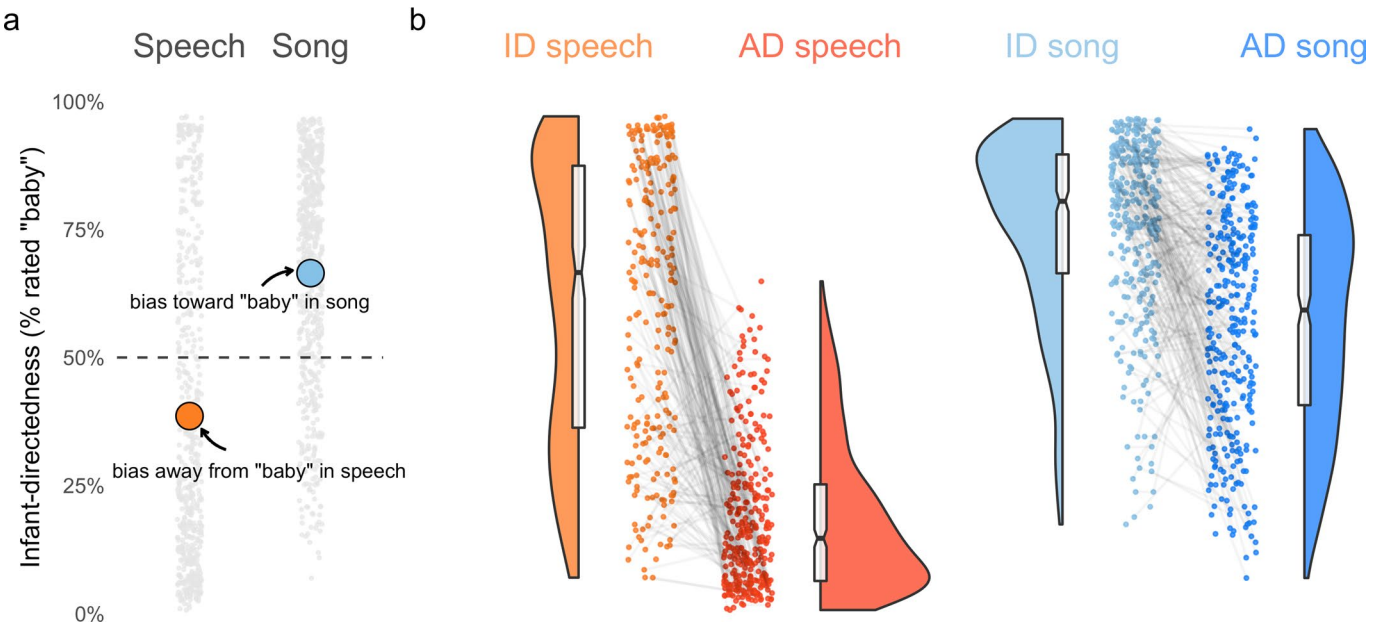
Extended Data Fig. 1 | Variation across societies of infant-directed alterations. Estimated differences between infant-directed and adult-directed vocalizations, for acoustic feature, in each fieldsite (corresponding with the doughnut plots in Fig. 2). The estimates are derived from the random-effect components of the mixed-effects model reported in the main text. Cells of the table are shaded to facilitate the visibility of corpus-wide consistency (or inconsistency): redder cells represent features where infant-directed vocalizations have higher estimates than adult-directed vocalizations and bluer cells represent features with the reverse pattern. Within speech and song, acoustic features are ordered by their degree of cross-cultural regularity; some features showed the same direction of effect in all 21 societies (for example, for speech, median pitch and pitch variability), whereas others were more variable.



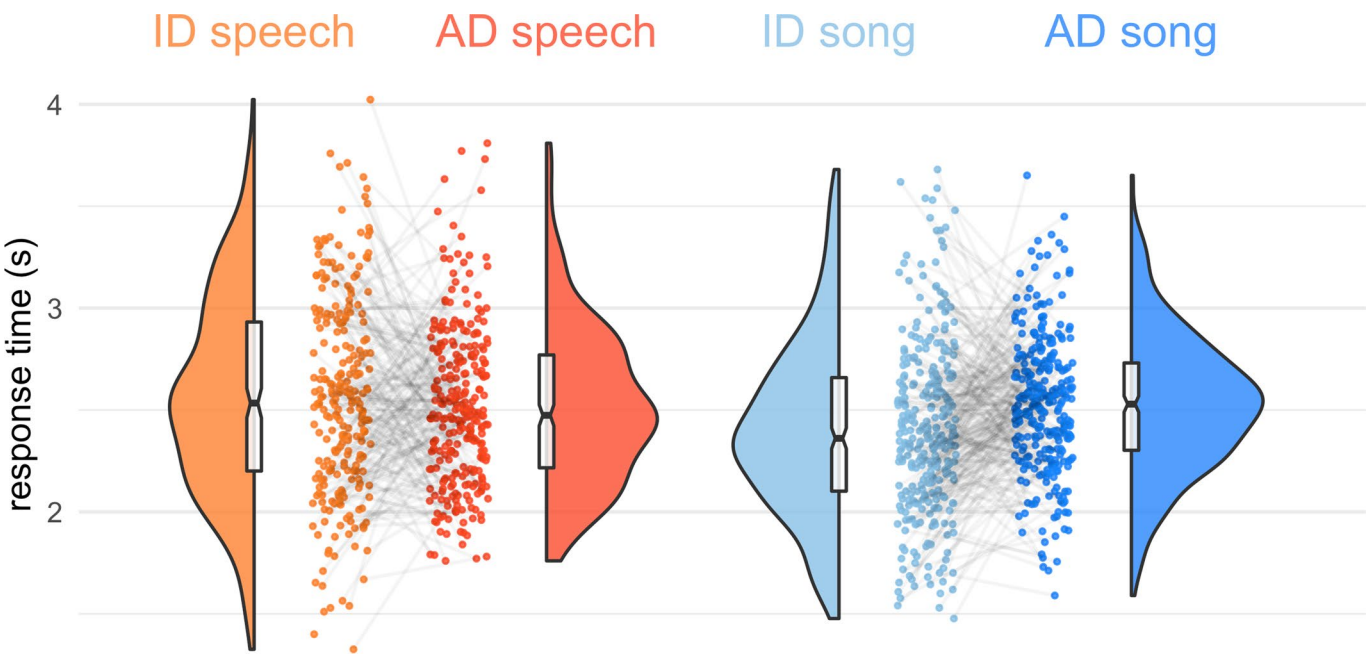
Extended Data Fig. 2 | Principal-components analysis of acoustic features. As an alternative approach to the acoustics data, we ran a principal-components analysis on the full 94 acoustic variables, to test whether an unsupervised method also yielded opposing trends in acoustic features across the different vocalization types. It did. The first three components explained 39% of total variability in the acoustic features. Moreover, the clearest differences between vocalization types accorded with the LASSO and mixed-effects modelling (Figs. 1b and 2). The first principal component most strongly differentiated speech and song, overall; the second most strongly differentiated infant-directed song from adult-directed song; and the third most strongly differentiated infant-directed speech from adult-directed speech. The violins indicate kernel density estimations and the boxplots represent the medians (centres), interquartile ranges (bounds of boxes) and 1.5 × IQR (whiskers). Significance values are computed via two-sided Wilcoxon signed-rank tests (n = 1,570 recordings); *p < 0.05, **p < 0.01, ***p < 0.001. Feature loadings are in Supplementary Table 7.



Extended Data Fig. 3 | Screenshots from the naive listener experiment. On each trial, participants heard a randomly selected vocalization from the corpus and were asked to quickly guess to whom the vocalization was directed: an adult or a baby. The experiment used large emoji and was designed to display comparably on desktop computers (**a**) or tablets/smartphones (**b**).



Extended Data Fig. 4 | Response biases in the naive listener experiment. a, Listeners showed reliable biases: regardless of whether a vocalization was infant- or adult-directed, the listeners gave speech recordings substantially fewer "baby" responses than expected by chance, and gave song recordings substantially more "baby" responses. The grey points represent average ratings for each of the recordings in the corpus that were used in the experiment (after exclusions, $n = 1{,}138$ recordings from the corpus of 1,615), split by speech and song; the orange and blue points indicate the means of each vocalization type; and the horizontal dashed line represents hypothetical chance level of 50%. **b**, Despite the response biases, within speech and song, the raw data nevertheless showed clear differences between infant-directed and adult-directed vocalizations, that is, by comparing infant-directedness scores within the same voice, across infant-directed and adult-directed vocalizations (visible here in the steep negative slopes of the grey lines). The main text results report only d' statistics for these data, for simplicity, but the main effects are nonetheless visible here in the raw data. The points indicate average ratings for each recording; the grey lines connecting the points indicate the pairs of vocalizations produced by the same voice; the half-violins are kernel density estimations; the boxplots represent the medians, interquartile ranges and 95% confidence intervals (indicated by the notches); and the horizontal dashed lines indicate the response bias levels (from a).



Extended Data Fig. 5 | Response-time analysis of naive listener experiment. We recorded the response times of participants in their mobile or desktop browsers, using jsPsych (see Methods), and asked whether, when responding correctly, participants more rapidly detected infant-directedness in speech or song. They did not: a mixed-effects regression predicting the difference in response time between infant-directed and adult-directed vocalizations (within speech or song), adjusting hierarchically for fieldsite and world region, yielded no significant differences (ps > .05 from two-sided linear combination tests; no adjustments made for multiple comparisons). The grey points represent average ratings for each of the recordings in the corpus that were used in the experiment (after exclusions, $n = 1{,}138$ recordings from the corpus of 1,615), split by speech and song; the grey lines connecting the points indicate the pairs of vocalizations produced by the same participant; the half-violins are kernel density estimations; and the boxplots represent the medians, interquartile ranges and 95% confidence intervals (indicated by the notches).